NARRATIVE

Navigating AI in Real-time for Reactive and Immersive Video-game Experiences

Andreas Jin Sum, Rune Korsgaard Ludvigsen, Nicolai Lind Bouquin

AALBORG UNIVERSITY

AALBORG UNIVERSITY
STUDENT REPORT

Department of Architecture, Design & Media Technology
Aalborg University
http://www.aau.dk

Narrative
NARRATIVE: Navigating AI in Real-time for Reactive and Immersive Video-game Experiences

Theme:
Generative AI Dialogue in Narrative-Driven Experiences

Project Period:
Spring 2024

Project Group:
Master Thesis

Participant(s):
Andreas Jin Sum
Rune Korsgaard Ludvigsen
Nicolai Lind Bouquin

Supervisor(s):
Henrik Schønau Fog

Copies: 1

Page Numbers: 34

Date of Completion:
May 23, 2024

Abstract:

Machine-learning and Artificial intelligence have risen in popularity in recent years, and significantly impacting the realm of video games. As chatbots become increasingly sophisticated and with more prevalent integration into interactive storytelling. Traditional multiple-choice dialogue systems can often be tedious and lead to players to skip through the conversation, missing out on potential narrative elements. This paper set out to explore the effect using an AI-chatbot would have on the player experience compared to traditional NPC dialogue systems. Through literature and state-of-the-art, the paper demonstrates how modern approach to narrative structure is seen in mainstream media and reflected for its incorporation with AI. A narrative experience was created in an AB test. A total of (N=19) participants took part in the study. A modified version of the Toronto Empathy Questionnaire was used to measure player responses in both experimental and control groups, revealing no significant difference in level of empathy. Using EMG as a form of verification has been previously proven to be effective, however, the results of the EMG in this study was found to be lacking, leaving a uncertainty of the usefulness of the readings. However, the study found a significant difference in the desire for continuation between the two groups, suggesting that AI NPCs may prove to enhance player experience.

Contents

1 Introduction

Machine-learning and trained Artificial intelligence (AI) have grown in popularity in recent years, as have their applications within video games. *Counter-Strike: Global Offensive* (CS:GO) (2012) is a game that utilises deep learning in their anti-cheat software. This was done by recording all online matches and whenever a player is reported for cheating by another player, these recordings are then sent to a system called Overwatch, in which other, external players will have to determine whether or not the accused player has cheated. In this manner, Valve, the developer and publisher of CS:GO, have been able to train their anti-cheat software to recognise cheaters in the game automatically, as covered at the Game Developers Conference in 2018 by John McDonald (GDC, 2020).
In 2018, OpenAI developed a bot that went on to defeat the world champions in the game *Dota 2* (2013). This was achived by utilizing reinforcement learning in a self-play environment, where the bot trained for 10 months (OpenAI, 2021).
In the game *Tomb Raider: Underworld* (2008), Machine Learning was used to analyze player behavior and create clusters of player profiles (Sifa et al., 2013). As chatbot AI has- and continues to improve, and new AIs constantly being released, it was only natural for them to appear in interactive storytelling as well. AI such as KoboldAI and AI Dungeon allow anyone to create their own text adventures by creating a story through interaction with the player. Inworld AI is an AI service that allows the designer shape the personality and background of the AI agent, as well as manipulation of responses in predefined and specified scenarios, providing a larger narrative control.

Non-Player Characters (NPCs) in games have significant influence on how the narrative and game is received (Inworld.Ai, 2023). Traditional hard-coded NPCs, can often leave a lot to be desired, with features like repetitive dialogue, lack of awareness of their surrounding environment and player actions. A self conducted study released by Inworld AI states that 28% of players dislikes the repetitiveness of traditional NPCs, and 22% dislike the lack of adaptation from said NPCs(Inworld.Ai, 2023).
According to the whitepaper released by Inworld AI, 99% of 1002 U.S based gamers believes that AI NPC's will improve gameplay (Inworld.Ai, 2023). Furthermore, 78% believe they would spend more time playing said games if they implemented adaptive AI(Inworld.Ai, 2023).
Further on in the paper, it is established that poor-quality NPCs often gets *trolled*, potentially breaking engagement with the narrative(Inworld.Ai, 2023). While these statements are intriguing and holds potential, the lack of proper empirical data to support this is concerning. This study would like to explore this notion of poor-quality NPCs getting treated worse, if empathy plays a role within this scenario and the overall player engagement.

A study explored the effect of an AI NPC with an emotional filter enabled, which showed that the participants were more inclined to engage and spend more time conversing with the NPCFraser, Papaioannou, and Lemon, 2018. Another paper proposed mixed initiative dialogue to help immersion. Using multiple choice dialogue, some of the options would lead the conversation while others simple served as basic responses. Based on the player inputs, the NPC would pick either leading or non-leading reponse. This was in an effort to discover a better and more dynamic solution to normal multiple choice dialogue, in which the players often feel forced in a specific conversational direction(Takahashi, Tanaka, & Oka, 2018).

For narrative purposes and comprehension, player concentration, engagement and empathy are crucial for player immersion within narrative experiences. As shown by Qin, these dimensions, among others were highlighted in created an immersive narrative experience(Qin, Patrick Rau, & Salvendy, 2009). With the possibilities of AI and Inworld AI's report, it should be a significant improvement to immersion in narrative experiences. Should AI NPCs be able to convey empathy and engage the player further than traditionally predefined NPCs, it should serve as a powerful tool for creating said experiences.

Extended research on the term *serious games*, also shows how video games and narrative experiences can be used for more than just entertainment(Hamari et al., 2016)(Naul & Liu, 2020). This includes, but is not limited to educational, informational and training. By being able to test the limits of AI, the potential of narrative experiences could see further improvements.

Understanding the impact that AI has on empathy can assist guiding the way for teaching communication and social ques to socially impaired people, as demonstrated in *Social Fringe Dwellers: Can chat-bots combat bullies to improve participation for children with autism?*(Ireland, Bradford, & Farr-Wharton, 2018). Training empathy and social interaction can further help with better integration into society. Children with autism tend to gravitate towards social and communication technologies, such as smartphones and tablets(Ireland, Bradford, & Farr-Wharton, 2018). Video games have been shown to increase the ability to recognise emotions and handle bullies(Ireland, Bradford, &

Farr-Wharton, 2018). Therefore, a better understanding of the effects of AI in video games could prove beneficial for future research.

With the the evolution of AI, it seem to hold the potential for improving a plethora of fields that media can provide. Not only does it offer a possibility to improve the overall enjoyment and engagement of a narrative, for entertainment purposes. It also holds promise in understanding empathy and social training for socially impaired people. Further more, this could extend even further into other realms of areas, such as personalized coaching or teachings. However, this is all but potential if not explored thoroughly. As such, the propose of conducting a research based on measuring the effects of incorporating AI NPC compared to predefined NPCs is essential. Both in regards to the overall engagement as well as the empathetic relation to NPCs. Considering the aforementioned focus areas, a research question established:

How and to what extent does AI NPCs affect empathy and player engagement compared to traditional NPCs.

2 Previous Research

2.1 AI in Interactive Digital Narrative

The exploration of generating interactive digital experiences through the utilization of AI and machine learning has been subject to research for many years, to push what is possible within its domain (Guzdial & Riedl, 2016)(Jain et al., 2020)(Mateas & Stern, 2003). The persuasive power of interactive media such as games have also been shown to be a powerful tool to convey ideas and reflection through the power of procedural rhetoric (Bogost, 2010). Generative AI has the possibility to aid this concept due to the possibility of the world reacting to the players action directly. This support the notion that players feel more engaged and invested if they perceive themselves as active participants in the narrative (Dow, 2007).

Earlier efforts centered around dynamic storytelling that is utilizing AI and tried to circumvent the traditional rule-based- and scripted interactions was *Façade* by Micheal Mateas and Andrew Stern (Mateas & Stern, 2003). The experience, from 2005, has the user enter a drama in a virtual world, using natural language understanding (NLU) & natural language processing (NLP), and without giving any directions. The system receives surface text input from the user and NLP is used to try analyze emotional cues and determine its meaning, for NLU to map the intent into acts and dynamically adjust character behaviors to maintain coherence (Mateas & Stern, 2003). The achievements of the system is its ability to create a sense of agency for the player, where their decisions directly influence the progression and outcome of the narrative, successfully simulating social interaction dynamics, blurring the line between story and game due to immersion (Knickmeyer & Mateas, 2005). Despite the its potential, *Façade* was held back due to technological advancements, as the drama was limited by authored events, pre-recorded voice acting, and the capabilities of the systems used at the time. This has led to the continued inspiration of the development of interactive narratives in video games and other digital media. Systems that leverage deep learning have been able to generate narratives that respond, seemingly intelligently, directly to a user's input. With language models such as Generative Pre-trained Transformer (GPT-2) showning promise in performance on various tasks, such as reasoning and sentiment analysis, without supervised data (Radford et al., 2019). Incorporation of these models has been used to create interactive text-based stories, with services such as AI Dungeon (Latitude-Inc., 2019). However, these models were still showing their limits when it came to consistency and coherence during longer sessions (Ammanabrolu et al., 2021)(Rashkin et al., 2020).

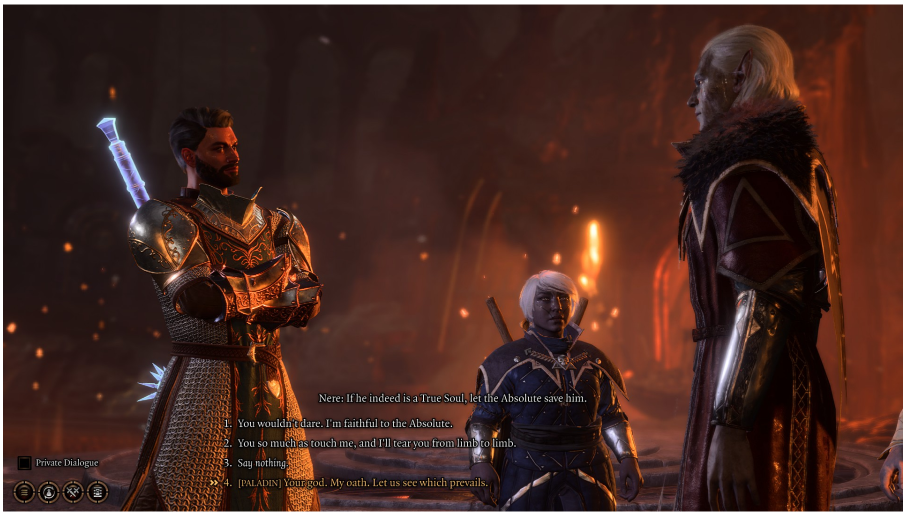
Not only does AI-driven narrative system contribute to an enhanced narrative experience, it also extends its application into other domains; educational simulations, therapeutic and artistic installations are but a few areas in which the leverage of AI to create immersive, engaging and personalized experience can creating exploration, and reflection (Cassell et al., 1999)(Cassell, 2001).

2.2 Risk of Generative AI

The capabilities of generative language models has been explored by integrating methods from psychology and philosophy (Sobieszek & Price, 2022). Despite their proficiency in generating plausible text, study reveals that statistical language generators lack consistency of reality, suggesting potential consequences for information reliability and trust (Krügel, Ostermaier, & Uhl, 2023) (Sobieszek & Price, 2022). While there is a lack in semantic ability and potential the risk of generating falsehood, as shown in Turing Test of GPT-3, further improvements was still shown compared to earlier models (Sobieszek & Price, 2022). Correspondingly, while models such as ChatGPT for story generation have been found to have inherent biases in the model's output, with a notable tendency to favor positive themes and adhere to restriction training, the potential for malicious input injection from users poses significant risks. An issue that highlights the dire need for solutions for optimized content filtering in order to mitigate otherwise harmful outcomes (Taveekitworachai et al., 2023). As such, these issues could further influence a narrative experiences negatively, as agent personality and coherence could break or risk generating harmful output and should therefore be considered during development.

2.3 Traditional NPCs

This paper makes use of the term *traditional NPC* which, for the purpose of this study, can be considered to be any system not utilizing dynamic generation or otherwise incorporate AI in direct interaction with the user. This study does not consider the use of AI beyond personalized experiences feedback, such as animation or path-finding. This traditional form of NPCs often offer dialogue system in which the player is able to choose between predefined conversational options, defined through a dialogue tree (Lessard, 2016). Most often this system is seen in state of the art games, such as Baldur's Gate 3, as seen on figure 1 (Larian Studios, 2023). Static dialogue systems can make interactions feel repetitive and unengaging (Strong & Mateas, 2008). These dialogue trees force the narrative onto the player, and while positive for guiding the story forward, players feel that they lack proper agency (Willerton, 2000).



Nere: If he indeed is a True Soul, let the Absolute save him.

1. You wouldn't dare. I'm faithful to the Absolute.
2. You so much as touch me, and I'll tear you from limb to limb.
3. *Say nothing.*
4. [PALADIN] Your god. My oath. Let us see which prevails.

□ Private Dialogue

Figure 1: Dialogue system used in Baldur's Gate 3 (Larian Studios, 2023)

While traditional dialogue systems can differentiate between experiences, they all share a common adherence to a narrative structure and is incapable of diverting off the set course(Ip, 2011)(Ryan, 1991). Players who are less engaged with an experience, also tend to skip dialogue for a merit of reasons, including being bored, interrupted or otherwise distracted(Mori & Miyake, 2022). Therefore, looking into player engagement in connection with narrative experience is crucial.

2.4 Interactive Narrative Structure

Shaping the user experience within an interactive narrative environment, the narrative structure plays a fundamental role in defining the narrative discourse and the player's agency. Drawing from Marie-Laure Ryan's definition of narrative structures, multiple elements are imperative in developing an engaging narrative experience (Ryan, 1991)(Ryan, 2015). The framework within which the story unfolds, comprising the arrangement of the narrative elements, such as events, relationships, and the temporal timeline that shape the coherence and progression of the narrative (Ryan, 1991). By providing an overview and using the narrative structure, generative AI has the potential to expand narrative experiences, providing users with even higher levels of player agency and engagement in interactive storytelling environments. While Ryan proposes multiple models for narrative experiences, few stand out aptly for use in the context of generative AI and most traditional dialogue system.

2.4.1 The Complete Graph

Each node in this structure is connected to every other node, providing the player with unrestricted access to every available narrative element. This allows for nonlinear exploration and traversal within the experience, encountering different story branches and possible outcomes.

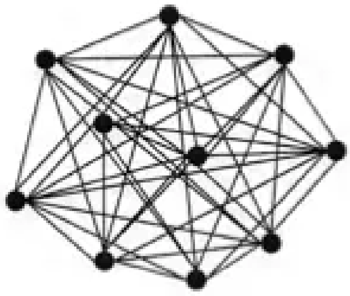


Figure 2: The complete graph structure as proposed by Ryan (Ryan, 2015)

2.4.2 The Tree

This hierarchical structure allows the player to branch out from a central point. Users progress through the narrative making choices that, in turn, determine the subsequent paths, leading to different narrative outcomes.

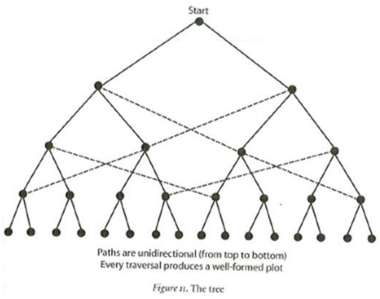

Start
Paths are unidirectional (from top to bottom)
Every traversal produces a well-formed plot
Figure 11. The tree

Figure 3: The Tree structure as proposed by Ryan (Ryan, 2015)

2.5 Traditional Narrative

An example can be seen in figure 4, of how these narrative structures are used in creating narrative systems. The figure shows the graph of a scenario within the state of the art narrative game Detroid Become Human, in which the user guides the narrative through a structure that most resembles with *The Tree* (Quantic Dream, 2018).

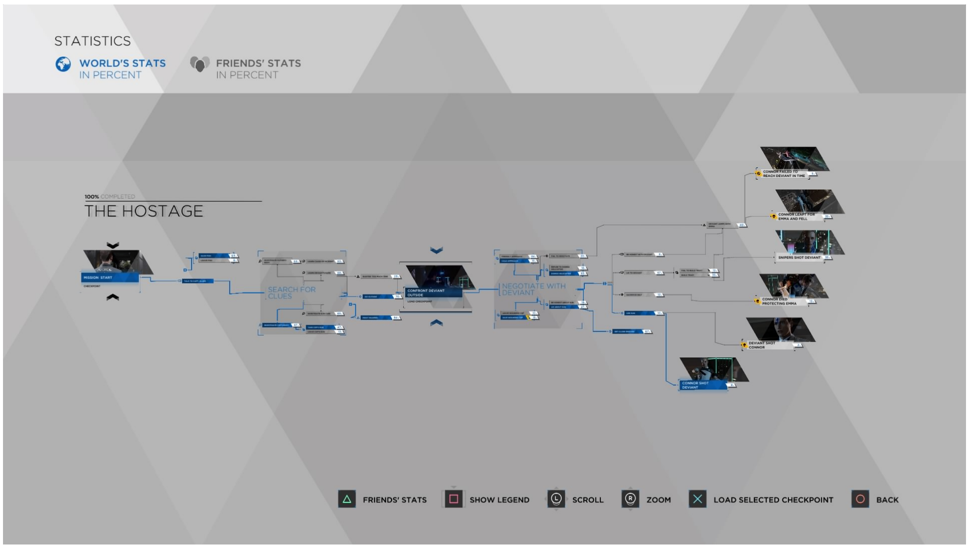


Figure 4: Screenshot of the of the narrative structure of a scenario in Detroit Become Human (Quantic Dream, 2018)

2.5.1 Generative AI and Narrative Structures

AI introduces new possibilities for increased emergent storytelling and gameplay, generative AI can dynamically generate narrative content based on user input, creating personalized narrative(Santiago III et al., 2023)(Ishii et al., 2019). While each narrative structure offers distinct avoidance for emergent storytelling, AI offers new approach to narrative experiences. Harnessing AI for the narrative generation, developers can create dynamic and adaptive narrative that is less constrained than pre-defined storytelling structures.

2.6 Player Engagement

Engagement is a way to determine the level of enjoyment a player experiences while interacting with a game or interactive story world, and can therefore be a good metric for measuring the success of enjoyment in a game (Lyons et al., 2014)(Chen et al., 2006). The relationship between engagement, narrative interest and immmersion in video games is complex, but several studies have examined this narrative and how the elements and player engagement enhance immersion for an overall improved experience(Qin, Patrick Rau, & Salvendy, 2009)(Hafner & Jansz, 2018).

One way to measure engagement in players is through continuation desire combined with narrative and user engagement, which cover the dimensions of; Attention Focus, Narrative Presence, Comprehension of Narrative, Emotional Engagement and Character Identification, Experimentation, and Disengagement causes. These dimensions can be used to help quantify the continuation desire within interactive storytelling through self-reported measures (Schoenau-Fog et al., 2013). Continuation Desire is a method for measuring player engagement in games and interactive media. The metric is based on the participants willingness to continue with the experience and a deeper investigation of the method categorises the different factors that affect the participants experience. The first factor is intrinsic and extrinsic objectives, the player and game sets goals that must be completed, this could be beating a specific boss, completing an achievement or the player willingly want to try out a specific interaction.
Activities describes the way players want to reach the objectives and varies greatly between games and type of players. Activities can be split into different categories: Solving, Sensing, Interfacing, Exploration, Experimentation, Creation, Destruction, Experiencing the Story, Experiencing the Characters, Communicating, Competing, and Socializing. Accomplishments occur when the player reaches a goal and is split into the categories achievements, Progression, and Completion. The last motivator is affect, which contains the emotional motivations the players might have while playing which can be positive affect, negative affect, and absorption (Schoenau-Fog, 2011b).

Evaluating through the use continuation desire can be achieved with the Engagement Sample Questionnaire (ESQ). ESQ is done in four parts, with the first part consisting of general demographic questions. The second part takes place before the experience and explores the participants expectations on the upcoming experience(Schoenau-Fog, 2011a). The third part of the ESQ takes place during runtime of the experience and explores the participants desire to continue with the experience at the moment of interruption and inquired as to why. The final part consists of a post experience questionnaire which explores the overall experience and once inquire for their desire to retry the experience (Schoenau-Fog, 2011a).

2.7 Empathy

Empathy lacks a universal definition, creating a larger issue in specific test variables(Neumann et al., 2014). However, *Measures of Empathy: Self-Report, Behavioral, and Neuroscientific Approaches* tries to combine the most commonly used definitions (Neumann et al., 2014):

> *Empathy involves an inductive affective (feeling) and cognitive evaluative (knowing) process that allows the individual to vicariously experience the feelings and understand the given situation of another. Its presence or absence is related to autonomic nervous system activity and overt behaviors that are augmented by affective intensity and cognitive accuracy. Further, empathy is a fundamental emotional and motivational component that facilitates sympathy and prosocial behavior (responding compassionately)*

Using this definition of empathy, they explored different kind of methods for measuring empathy. This includes self-report and behavioral methods consisting of questionnaires, and most Neuroscientific methods, often requiring intrusive or expensive equipment. With this in mind, most neuroscientific methods were deemed out of scope for this project, with the exception of electroencephalogram (EEG) and facial electromyography (EMG), with EMG being the less intrusive, focusing on measuring the electric potential produced by skeletal muscles. To identify the need for reliable and non-intrusive methods, self-report questionnaires were explored. Among these, the Toronto Empathy Questionnaire (TEQ) is a recent and widely applicable method that requires minimal altercations to encompass this project(Neumann et al., 2014).

2.7.1 EMG

Facial electromyography (EMG) captures muscle contraction that is below the visual threshold, it is proposed that this supports empathic responses(Neumann et al., 2014). In order to capture facial emotions, electrodes are attached to the skin over specific muscle regions that correspond to the desired emotions. The most frequent found ones are the *corrugator supercilli, zygmaticus major, lateral frontalis, medial frontalis, levator labii superioris, orbicularis oculi and masseter*(Neumann et al., 2014). The measurements focus on the signal's magnitude and which muscle groups are involved. However, due to the nature of the process, there is a high potential bias due to external disturbances,

such as sound as well as the face electrodes may cause people to become more aware of their own facial expressions, which could lead to skewed data(Neumann et al., 2014).
In his chapter in the *handbook of emotions*, Matsumoto presents a table that use Darwins descriptions of emotions and correlates them with the action unit(AU) described in the Facial Action Coding System (FACS)(Matsumoto et al., 2010). Furthermore, each AU from the FACS can be interpreted and reduced to specific muscles within the face (Jiang et al., 2015). A combination can be seen in figure 5.

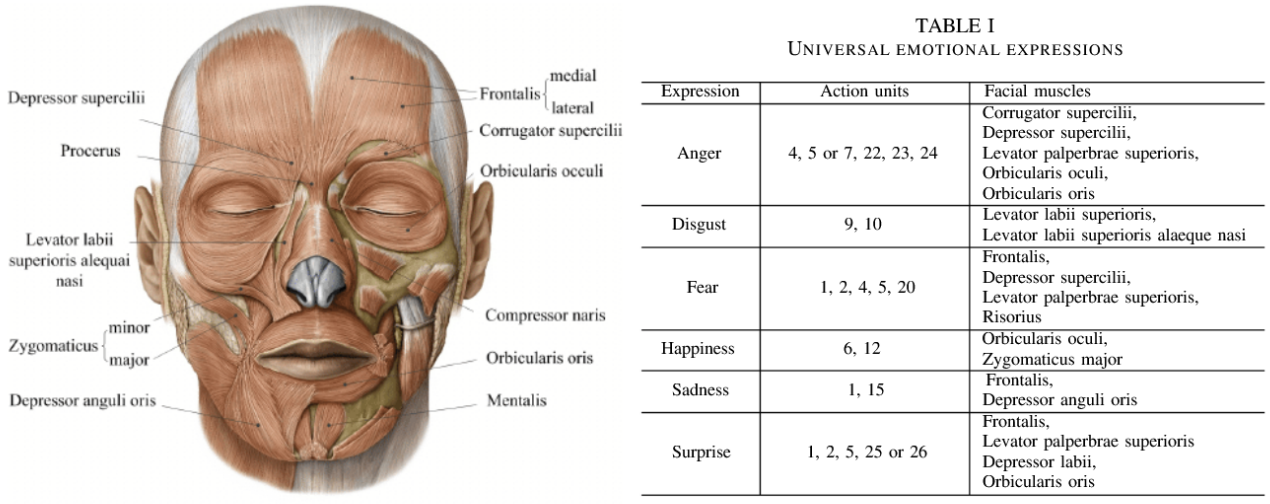


TABLE I
UNIVERSAL EMOTIONAL EXPRESSIONS

Expression	Action units	Facial muscles
Anger	4, 5 or 7, 22, 23, 24	Corrugator supercilii, Depressor supercilii, Levator palperbrae superioris, Orbicularis oculi, Orbicularis oris
Disgust	9, 10	Levator labii superioris, Levator labii superioris alaeque nasi
Fear	1, 2, 4, 5, 20	Frontalis, Depressor supercilii, Levator palperbrae superioris, Risorius
Happiness	6, 12	Orbicularis oculi, Zygomaticus major
Sadness	1, 15	Frontalis, Depressor anguli oris
Surprise	1, 2, 5, 25 or 26	Frontalis, Levator palperbrae superioris Depressor labii, Orbicularis oris

Figure 5: Shows the different muscle groups and their location as described in (Jiang et al., 2015)

A less intrusive method would be utilizing computer vision (Yang & Bhanu, 2012) (Carvalhais & Magalhães, 2018), with further using machine learning to recognise emotions and reducing the emotions to core emotions (Carvalhais & Magalhães, 2018) (LokeshNaik et al., 2023),(Bakariya, Singh, & Singh, 2024). however, through testing an accuracy was deemed too low. (Further described in Appendix B)

2.7.2 TEQ

The TEQ, was created using an exploratory factor analysis to establish a unidimensional factor for the questionnaire, it incorporates multiple of the self-report questionnaires and scales as discussed in *Measures of Empathy: Self-Report, Behavioral, and Neuroscientific Approaches* (Spreng et al., 2009). The questionnaire has been through validation three studies demonstrating both its reliability and validation. For this project, modifications are needed as TEQ is a broad generalized questionnaire, where as for this project the empathy is not targeted in general but at specific NPCs. The TEQ has be retested and validated more recent in multiple languages, which further solidifies TEQ validity (Janelt et al., 2023) (Xu et al., 2020) (Totan, Tayfun, & Sapmaz, 2012) (Kourmousi et al., 2017).

2.8 Summary of Prior Findings and Hypothesis Formulation

Drawing on the related literature, it becomes evident that while AI have the potential to affect the user experience, there is a lack of empirical data comparing their effectiveness against traditional NPCs in enhancing the overall engagement and empathy is sparse. With this in mind, this study will aim to fill these gaps by further investigating the following hypothesis through testing:

Null Hypotheses (H0):

1. $H_{0,1}$: AI NPCs do not significantly increase empathy in players compared to traditional NPCs.
2. $H_{0,2}$: AI NPCs do not significantly enhance player engagement compared to traditional NPCs.

Alternative Hypotheses (H1):

1. $H_{1,1}$: AI NPCs significantly increase empathy in players compared to traditional NPCs.
2. $H_{1,2}$: AI NPCs significantly enhance player engagement compared to traditional NPCs.

3 Design

This section will go over the methods that are planned to be used in correlation with the study. It outlines the intended framework utilized to ensure a comprehensive and systematic examination of the research. The methods chosen were carefully chosen with the study's objective and previous research, in order to best ensure reliability and replicability of the results. The section will also cover the data collection method and sampling strategy to offer transparency of our finding.

3.1 TEQ

TEQ is based on a generalized level of empathy not taking into account the personality and level of familiarity with the subject, as such for this experiment the questionnaire was altered to focus on a specific NPC. The changes consists of two types of changes; rewriting and removal of items. The removed questions items are 10, 11 and 15, as they were deemed not of relevance or too complex for rewriting for this study, which could lead to confusion. The remaining items were rewritten, with the goal of changing any reference from someone/anyone to specifically point to the NPC in question. As the a new version of the TEQ has been created it needs to further be re-validated as such using the test participants for the final test, the questionnaire will be validated for potential future use.

3.2 EMG

The experience will focus on specific emotions to minimize amount of electrodes needed on the participant, as such sadness was chosen, based on muscle groups sadness was decided as the main emotion to test for, however happiness could have been equally as useful, although a fun and happy gameplay loop will reduce the overall difference between base and emotion inducing event (A sad event will have a higher difference compared to a happy event, if the gameplay loop too is happy). As such The electrodes are to be placed on the Frontalis, which is located in the forehead going towards the eyebrow and the Depressor anguli oris, located at the chin going towards the cheekbone as seen in figure 5.

3.3 Test Procedure

In order to best achieve proper testing and cohesion, this section provides detailed instructions for conducting the test for the study, including the different sequences of activities and steps that must be followed, for both versions of the platform. It describes the scope, required equipment and the overall procedure.

3.3.1 Purpose and Scope

This test procedure outlines the steps for a lab study examining the relationship between AI, player agency and emotional response, in a narrative-driven interactive experience. The study consists of two versions; an experimental version, where participants engage with AI-driven NPC, and a control version where they follow a predefined narrative with multiple choice dialogue. The expected duration for each participant is approximately 15 minutes in which 5 minutes are dedicated to the playthrough. The minimum expected participation count is 20, with 10 participants for each version.

3.3.2 Test Responsibilities

The test setup involves two team members coordinating the process: a Test Conductor and an Assistant. Both the Test Conductor and the Assistant are responsible for maintaining a safe test environment and ensuring the participant feels comfortable throughout the process. Together, they create a seamless and efficient testing experience.

Test Conductor:

- Greets the participant, explains the test process, and provides instructions.
- Guides the participant through the preliminary steps, including filling out information and using wet wipes to clean the sensor areas.
- Places and adjusts the EMG sensors on the participant's face.
- Gives gameplay instructions and monitors the participant during the game session.

- Conducts the feedback session after gameplay, asking for a 1-to-10 rating on the participant's interest in continuing the game, and initiates the final questionnaire.
- Removes the sensors after the test.

Assistant:

- Records data from the EMG sensors during the test session.
- Prepares the EMG equipment before each test.
- Assists with sensors and other technical tasks, as needed.
- Ensures all equipment functions properly throughout the test and addresses any technical issues.
- Coordinates the collection and organization of test data results after the session.

3.3.3 Equipment and Materials:

a list of all required equipment and materials for the test, ensuring readiness and proper setup are as follows

- Computers with the interactive experience (Experimental and Control version) as well as the questionnaire.
- EMG sensors (e.g., Trigno Mini) - including single use adhesives for electrodes.
- A second computer running necessary software for recording EMG data (EMGWorks aquisition).
- Disposable cleaning wipes for sanitizing equipment and participant's contact area, in accordance to specified instructions provided by Trigno Mini Manual.

3.3.4 Preparation and Setup

Preparations and execution are the same for both the experimental and control version of the platforms, with the only difference being the chosen version executed.

3.3.5 Briefing of Participants

Briefing outlines the initial briefing process for participants, ensuring they understand the study's objectives, procedures, and their role in the test.

- The participants are welcomed and provided with proper information about the test process and the consent form. (Full script and walk-through can be found in C).
- Ensure that they understand the use of the EMG sensor.

3.3.6 Placement of Electrodes Trigno Mini:

Proper placement of electrodes are essential for ensuring accurate measurement and monitoring of participant responses.

- The area of contact is cleaned with wet wipes.
- Attach the EMG sensors to the Frontalis, Depressor Angulis Oris and common reference points - ensuring proper contact.

3.3.7 The Experience Playthrough:

The step of initiating and having the participants play.

- Brief participant on control scheme.
- Start the narrative-driven experience.
- Observe participants' interactions and note key moments.

3.3.8 End Experience:

After the experience has concluded, the conductor stops the participants, and query their desire to continue (on a 1-10 scale). Before moving unto the final questionnaire, and conclusion to the test.

- Gently remove EMG sensors.
- Have Participants fill out questionnaires.
- Thank the participants for their time.

3.3.9 Cleanup/Disposal

After the test has concluded the steps for cleaning and disposing of materials, of single-use items can begin.

- Sanitize equipment after each participant.
- Dispose of any single-use materials like wipes or sensor pads.
- EMG Data is saved and exported.

3.4 Recruitment of Participants and Platform Assignment

Recruitment of participants will be drawn through convenience sampling, where individuals will be approached at the university of AAU Copenhagen. Here they will be invited to partake in the study and invited to the lab. If possible, participants will be scheduled into time slots to avoid overlap and ensure a streamlined test process.

Due to the structure created for this research, the study can be considered to be a between-subject design, in which participants will be randomly assigned to one of the two platforms; either the experimental group (A) or the control group (B). The list will be randomized in advanced with the expected sample size of at least 20 participants, with the goal of having even distribution of having 10 participants in each group.

3.5 Testing and Data analysis

In order to test the effect of empathy, emotional response and engagement, the collected data will undergo analysis. As such the plan for the data analysis for the questionnaire regarding empathy is as follows:

1. **Reliability of Questionnaire** - The initial step will be testing whether or not the modifications to the questionnaire still contains reliability. This can be achieved using Chronbach's Alpha.
2. **Normality** - Should the questionnaire prove to be reliable, the data should test for normality. Due to the potential of a smaller sample size and the data being quantitative numerical, Shapiro-Wilks test is an effective choice to test.
3. **Statistical Test** - Assuming the data follows normal distribution, an unpaired t-test will be performed in order to test the means between the independent testing groups. Should the data prove to not assume normality, a non-parametric test such as Mann-Whitney U Test will be executed instead, comparing the two groups.

After testing the responses from the questionnaire, the scoring of the continuation desire will go through a similar treatment in order to test the statistical difference in player engagement and desire to play:

1. **Normality** - Due to the potential of a smaller sample size and the data being quantitative numerical, Shapiro-Wilks test is an effective choice to test for normality.
2. **Statistical Test** - Assuming the data follows normal distribution, an unpaired t-test will be performed in order to test the means between the independent testing groups. Should the data prove to not assume normality, a non-parametric test such as Mann-Whitney U Test will be executed instead, comparing the two groups.

3.6 Analysis of EMG

EMG data can present as both positive and negative values, so in order normalize the data Root Mean Square(RMS) will be applied to it. To extract information from the EMG data it first needs to be cleaned, this will be done by removing samples outside the point of interest and reducing all participants to the same amount of samples. The collected data will contain multiple unnecessary columns which need to be dropped, until only the depressor RMS and the Frontalis RMS are left. Then the mean for each column will be calculated for each participant. The clean data will then be analyzed similarly to how the other data will be.

1. **Normality** - Due to the potential of a smaller sample size and the data being quantitative numerical, Shapiro-Wilks test is an effective choice to test for normality.

2. **Statistical Test** - Assuming the data follows normal distribution, an unpaired t-test will be performed in order to test the means between the independent testing groups. Should the data prove to not assume normality, a non-parametric test such as Mann-Whitney U Test will be executed instead, comparing the two groups.

3.7 Iterative Process

With chatbots being somewhat unpredictable and prone to hallucinations. Tests were conducted throughout the development process, with the focus on training the NPC. This test consisted of convenience sampling with an observer to focus on responses that were outside or contradicting to the trained prompts. Each test's responses were evaluated and action were taking to change training data, before another test was conducted. Initial iterations consisted of internal testing for the purpose of expedience and rough fine-tuning, where further in the process, in which the NPC was less prone to mistakes, external testing was conducted. This process allowed to minimize hallucinations and further minimize the impact which unforeseen hallucinations would have on the overarching experience.

The NPC's responses for the control version were iterated on in parrallel, using the newly trained NPC of each iteration. This allowed for responses to be as similar as possible between the versions. The responses were based on the interaction the test participants had with the NPC through the experimental iteration phase. Due to the specificity of some responses (referring to previous questions and responses), the responses went through a quality assurance, in which some where rewritten to be more generalized responses and not take into account what the experimental participant's prior input.

4 Implementation

This section provides a detailed overview and process of the implementation and development phases involved in creating the test platform. The platform make use of AI for NPCs for the experimental version and traditional NPC systems for the control. A quick description of the repositories, design process for user experience (UX) and usage of AI will be uncovered. The development process was guided by incorporating an iterative testing and feedback loop to refine the NPC behaviour and UX.

4.1 Chatbot - Bethy

For the actor in the scene an AI engine called Inworld was used to facilitate the creation and training of the chatbot. This allowed for a faster and more streamlined process of fine-tuning the behavior of the actor. The initial steps were to create a persona for this actor and for the sake of creating the best chance for a higher level of empathy. For this a caring and comforting persona was created in the form of an older lady named *Bethy*. Inworld allows for API use and an easy to integrate plugin for Unity. For creating an actor with a certain knowledge base, as well as trying to remove all *Hallucinations*, multiple iteration and test loops were performed changing knowledge location and phrasing as well as honing in the personality desired.

4.1.1 Goals

Apart from the standard interaction of chatting with the bot, a goals system is available. This allows for changes in personality to prompt specific responses to actions and triggers. In this project it was used to give tasks and thank the player for completing them using a mix of unity code and Inworld's own mark-up language (YAML editor). Most of the goals would be triggered through chat, in which an intent was used. Intent works by giving a series of training phrases on which to base the intent on. Inworld's system would then determine whether a responds is a close enough match to any of the training phrases and if so, completes the goal.
The goals can then be used to give instruction to the NPC in the form of traditional AI prompts such as *"Ask the player for help with p.task in a short manner"*.

4.2 Core Experience

The core framework of the experience shared between the two versions consist of the player walking around a café with the purpose of engaging in dialogue with *Bethy*. The player will through conversation learn more about *Bethy*and possibly receive tasks to help her around the café. For the purpose of the project, a total of five unique tasks can be given to the player; these tasks were hard coded and designed in order to avoid any unwanted issues and to ensure some consistency between the experiences. Through these tasks a simple gameplay loop was generated;

- Action: Talk to *Bethy*, complete her objective and return for more tasks.
- Goals: Complete the objectives from *Bethy*

A timer was set to initiate the conclusion of the experience. At the five minute mark, *Bethy* would close any ongoing conversation and the final pre-recorded event would take place.

4.2.1 Finale

The end of the experience would be mutual for both versions in order to minimize bias. The ending would initiate a phone call between *Bethy* and a disembodied murmur. The sequence was created with a pre-scripted and executed text-to-speech, and edited using Audacity. The intention was to create an emotional change within *Bethy*, with the goal of creating an emotional change by proxy for the user experiencing the sequence.

4.3 Experimental Version

For the experimental version, the experience would have full access to the Inworld chatbot of *Bethy*, offering an almost unrestricted conversation in which the player is able to express themselves as they desired.

To allow for more control the main task-giving system was implemented in Unity, as the Inworld goal and mutation system was found lacking for this specific task. This means that an intent is registered in inworld that sends a trigger to Unity that then invokes all function associated with a random tasks, including sending a package back to Inworld that triggers another prompt which says *"Ask the player for help with p.task in a short manner"*. the packages sent, includes the task in the form of p.task.

4.4 Control Version

The control is limited by the implemented dialogue options, which can be seen in figure 6. As the task system is already implemented through the experimental experience, it can be reused for the control, removing only the packages sent back and forth, this allow for

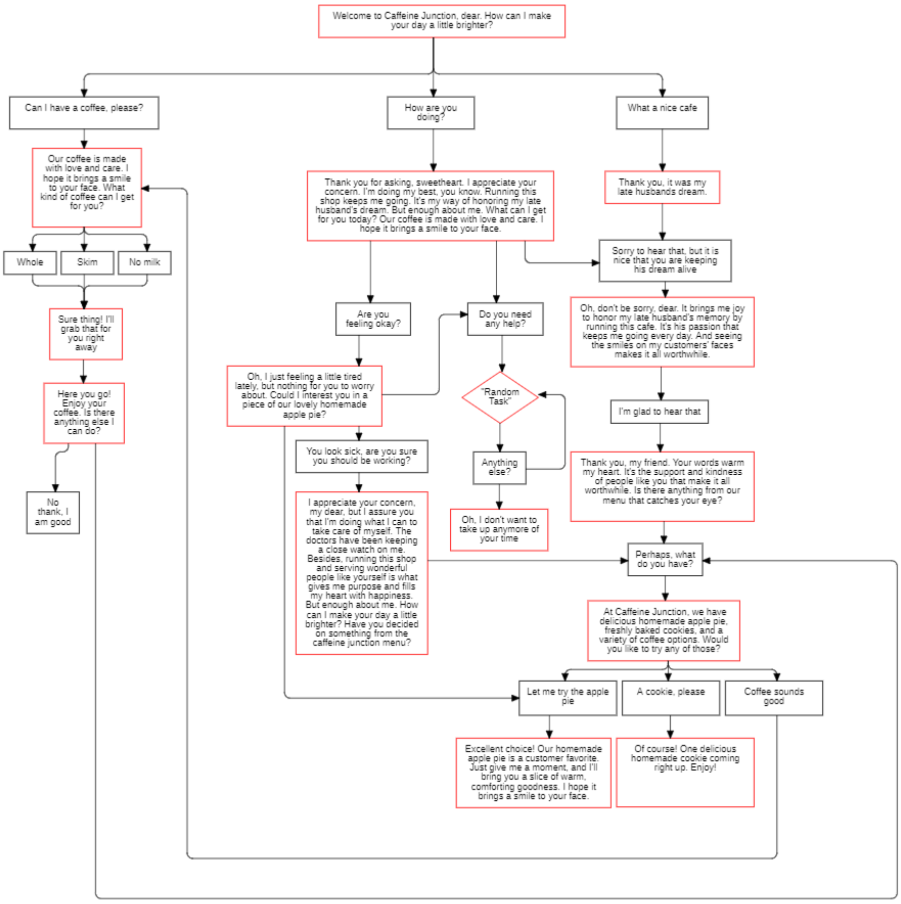

Welcome to Caffeine Junction, dear. How can I make your day a little brighter?
Can I have a coffee, please?
How are you doing?
What a nice cafe
Our coffee is made with love and care. I hope it brings a smile to your face. What kind of coffee can I get for you?
Thank you for asking, sweetheart. I appreciate your concern. I'm doing my best, you know. Running this shop keeps me going. It's my way of honoring my late husband's dream. But enough about me. What can I get for you today? Our coffee is made with love and care. I hope it brings a smile to your face.
Thank you, it was my late husbands dream.
Whole
Skim
No milk
Sorry to hear that, but it is nice that you are keeping his dream alive
Sure thing! I'll grab that for you right away
Are you feeling okay?
Do you need any help?
Oh, don't be sorry, dear. It brings me joy to honor my late husband's memory by running this cafe. It's his passion that keeps me going every day. And seeing the smiles on my customers' faces makes it all worthwhile.
Oh, I just feeling a little tired lately, but nothing for you to worry about. Could I interest you in a piece of our lovely homemade apple pie?
"Random Task"
Here you go! Enjoy your coffee. Is there anything else I can do?
I'm glad to hear that
You look sick, are you sure you should be working?
Anything else?
Thank you, my friend. Your words warm my heart. It's the support and kindness of people like you that make it all worthwhile. Is there anything from our menu that catches your eye?
No thank, I am good
I appreciate your concern, my dear, but I assure you that I'm doing what I can to take care of myself. The doctors have been keeping a close watch on me. Besides, running this shop and serving wonderful people like yourself is what gives me purpose and fills my heart with happiness. But enough about me. How can I make your day a little brighter? Have you decided on something from the caffeine junction menu?
Oh, I don't want to take up anymore of your time
Perhaps, what do you have?
At Caffeine Junction, we have delicious homemade apple pie, freshly baked cookies, and a variety of coffee options. Would you like to try any of those?
Let me try the apple pie
A cookie, please
Coffee sounds good
Excellent choice! Our homemade apple pie is a customer favorite. Just give me a moment, and I'll bring you a slice of warm, comforting goodness. I hope it brings a smile to your face.
Of course! One delicious homemade cookie coming right up. Enjoy!

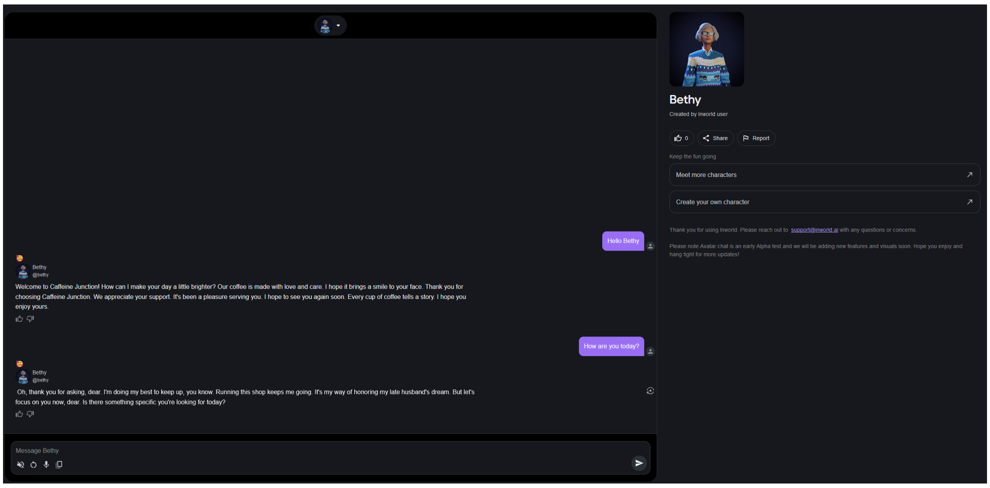
Figure 6: Dialogue tree for control version. (Larger version in Appendix A)

4.5 User Experience

Considerable effort was put into creating a smooth and effective user-friendly interface. The purpose was to integrate both the versions simultaneously in one overall theme to ensure consistency and avoid interfering with the user experience between the groups. As such, adopting an iterative development for the UI and general interaction was used.

The iterations of the UI and interactions were subjected to psuedo-usability testing by a selected group of participants. Based on their feedback changes were incorporated into the subsequent iterations. These tests however, were without formal data recording and lack proper scoring data, such as system usability scale (SUS). Despite this, the refinements to the project allowed for continual improvement and optimization of the UX design without the need for formal data collection and analysis at every stage of iteration.

The biggest difference in UX was the interaction between *Bethy* and the player. As the Experimental condition featured a completly free chat with *Bethy*, and had to facilitate the natural feel of conversation. The UI would come to reflect the original Inworld design, in which the conversation works like any regular modern messaging system, as seen in figure 7 & 8.



Hello Bethy

Bethy
@bethy

Welcome to Caffeine Junction! How can I make your day a little brighter? Our coffee is made with love and care. I hope it brings a smile to your face. Thank you for choosing Caffeine Junction. We appreciate your support. It's been a pleasure serving you. I hope to see you again soon. Every cup of coffee tells a story. I hope you enjoy yours.

How are you today?

Bethy
@bethy

Oh, thank you for asking, dear. I'm doing my best to keep up, you know. Running this shop keeps me going. It's my way of honoring my late husband's dream. But let's focus on you now, dear. Is there something specific you're looking for today?

Message Bethy

Bethy

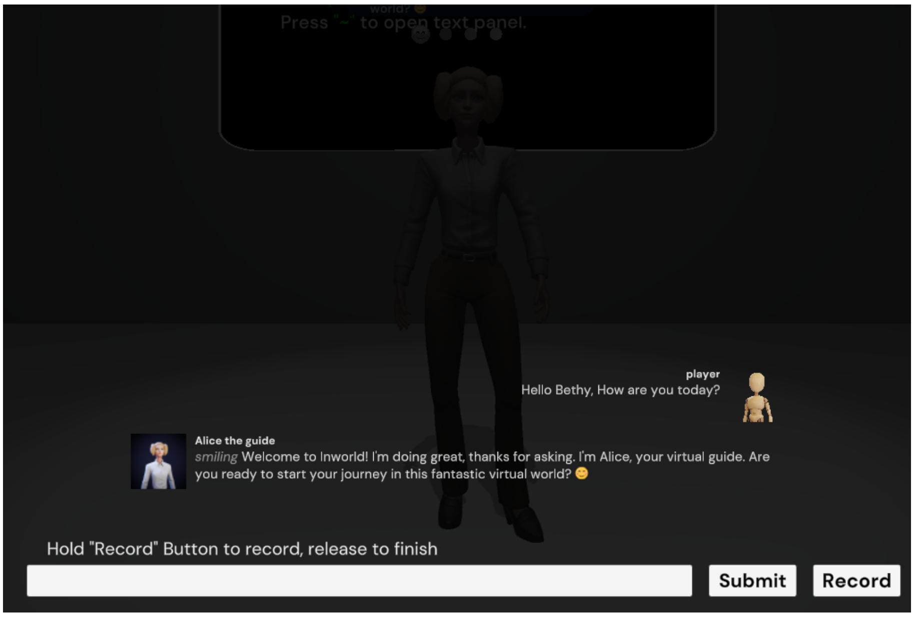
Created by Inworld user

Keep the fun going

Thank you for using Inworld. Please reach out to support@inworld.ai with any questions or concerns.

Please note Avatar chat is an early Alpha test and we will be adding new features and visuals soon. Hope you enjoy and hang tight for more updates!

Figure 7: User Interface seen at Inworld.ai



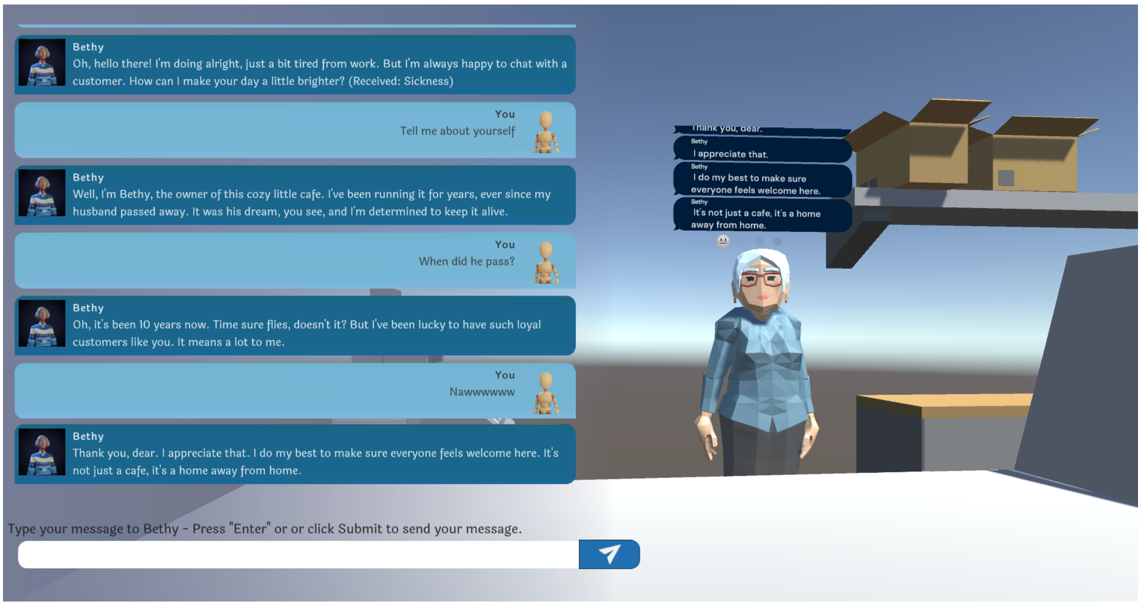
player
Hello Bethy, How are you today?

Alice the guide
smiling Welcome to Inworld! I'm doing great, thanks for asking. I'm Alice, your virtual guide. Are you ready to start your journey in this fantastic virtual world? 😊

Hold "Record" Button to record, release to finish

Figure 8: Inworld Unity package default UI.)

An early iteration, as seen in figure 9, tried to ease the cognitive load and maximise user autonomy by letting users express themselves. This included removing or editing unnecessary features and text. These changes included indicating clear separation between the chat bubbles and the possibility to look at the NPC talking, as well as other interface changes, like choice of font, color profile and implementation of iconography instead of text buttons.



Bethy
Oh, hello there! I'm doing alright, just a bit tired from work. But I'm always happy to chat with a customer. How can I make your day a little brighter? (Received: Sickness)

You
Tell me about yourself

Bethy
Well, I'm Bethy, the owner of this cozy little cafe. I've been running it for years, ever since my husband passed away. It was his dream, you see, and I'm determined to keep it alive.

You
When did he pass?

Bethy
Oh, it's been 10 years now. Time sure flies, doesn't it? But I've been lucky to have such loyal customers like you. It means a lot to me.

You
Nawwwwww

Bethy
Thank you, dear. I appreciate that. I do my best to make sure everyone feels welcome here. It's not just a cafe, it's a home away from home.

Thank you, dear.

Bethy
I appreciate that.

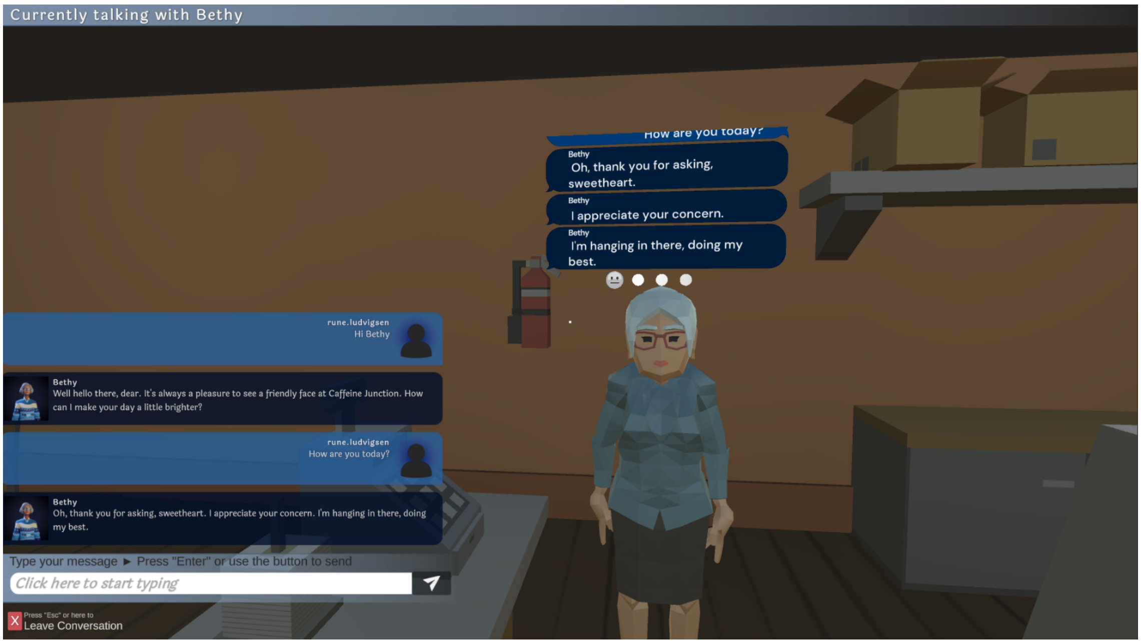
Bethy
I do my best to make sure everyone feels welcome here.

Bethy
It's not just a cafe, it's a home away from home.

Type your message to Bethy - Press "Enter" or or click Submit to send your message.

Figure 9: Early Iteration of the custom User Interface.

Based on feedback and testing of the design, minor adjustments and changes were constantly made until the final version, that proved to be highest liked and offered best usability. As seen if figure 10, this version had a much more compact and simplistic design, while offering high consistency and user-centeric design philosophy.



Currently talking with Bethy

How are you today?

Bethy
Oh, thank you for asking, sweetheart.

Bethy
I appreciate your concern.

Bethy
I'm hanging in there, doing my best.

rune.ludvigsen
Hi Bethy

Bethy
Well hello there, dear. It's always a pleasure to see a friendly face at Caffeine Junction. How can I make your day a little brighter?

rune.ludvigsen
How are you today?

Bethy
Oh, thank you for asking, sweetheart. I appreciate your concern. I'm hanging in there, doing my best.

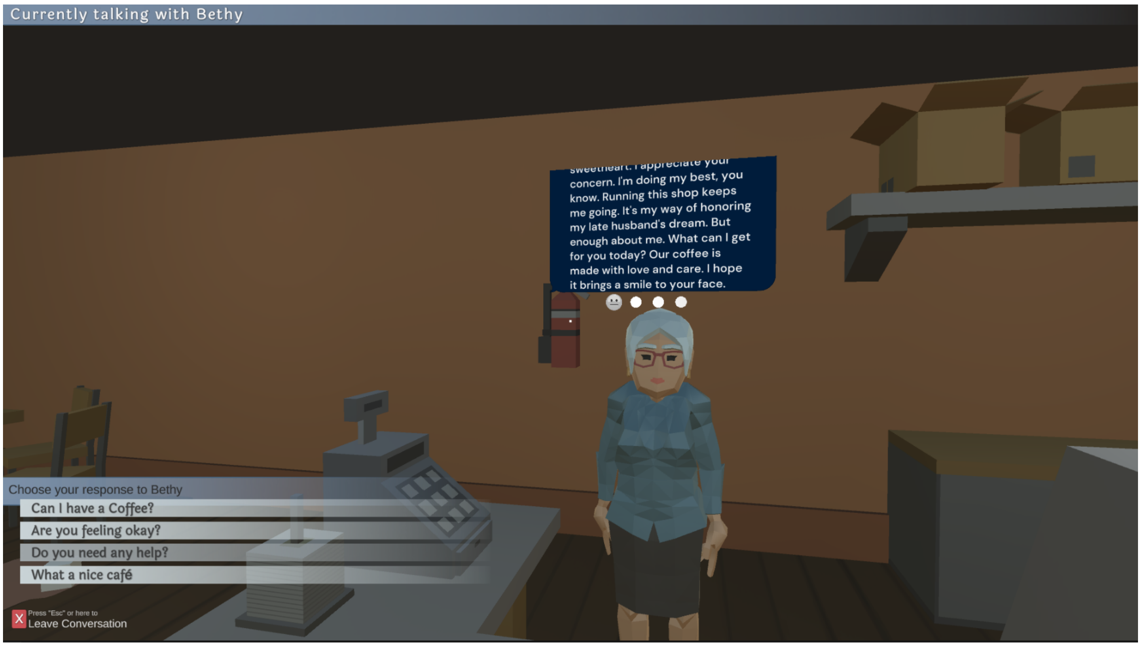
Type your message ► Press "Enter" or use the button to send

Click here to start typing

X Press "Esc" or here to
Leave Conversation

Figure 10: Final iteration of the chat User Interface.

For the control version, it was imperative to keep the consistency in theme and avoid overloading the user or otherwise make the version complex in comparison. As can be seen in figure 11, the UI adhere to the original UI, but implements a dialogue selection system, often seen in traditional video games. This system proved to be easy to navigate and efficient in use.



Currently talking with Bethy

sweetheart. I appreciate your concern. I'm doing my best, you know. Running this shop keeps me going. It's my way of honoring my late husband's dream. But enough about me. What can I get for you today? Our coffee is made with love and care. I hope it brings a smile to your face.

Choose your response to Bethy

- Can I have a Coffee?
- Are you feeling okay?
- Do you need any help?
- What a nice café

Press "Esc" or here to
Leave Conversation

Figure 11: User Interface for the control version, featuring dialogue options.

Another area of importance was the indicator for being able to interact within the game-world, shown in figure 12. This UI would likewise end up reflecting the theme and aesthetic of the UI while prompting guidance for the player in how to react, by incorporating iconography of a mouse with left button clicked.

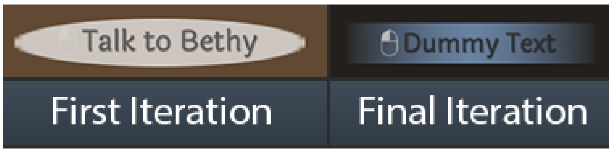


Figure 12: Initial and final iteration of the UI used for interacting within the experience.

5 Results

The data derived through the study compromised of the responses from (N=19) participants, consisting of 12 males, 6 females and one other. The mean age of the participants was 25 years ($SD = 2.36$). Participants were primarily recruited from the Aalborg University campus in Copenhagen through convenience sampling. Participants were approached and recruited through personal contact and brought to the lab for testing. The remaining participants were recruited through accessibility and availability, by leveraging personal connections and networks. The full data set for both groups, including each respondent can be found in Appendix D

5.1 Empathy

In order to answer the research question and related hypothesis, the data was treated and analyzed to find an answer whether or not the exposing to the experimental platform carried any change to the empathy of the user, compared to the control version.

5.1.1 Test for Reliability

For measuring the internal consistency of a scale or test, Chronbach's alpha provides insight to whether items in a scale are related, reflecting the reliability in measuring. As the questionnaire that was created was a modified version of the Toronto Empathy Questionnaire, this test was deemed necessary step for ensuring reliability of the measurement instrument. To test internal consistency, the data sets of both groups would be combined. Chronbach's Alpha can be calculated using the formula -

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_{Y_i}^2}{\sigma_X^2}\right)$$

where:

- N is the number of items.
- $\sigma_{Y_i}^2$ is the variance of item i.
- σ_X^2 is the variance of the total score formed by summing all items.

In Listing 1, is a presentation of the Python code for calculating Cronbach's alpha for the combined dataset. Initially filling missing values with column means and then reverse scores for specified items. Finally, the calculation of Cronbach's alpha using the pingouin package.

```
    alpha = pg.cronbach_alpha(data)
```

```
    reverse_items = [2, 4, 7, 10, 12]
    combined_df_filled = combined_df.fillna(combined_df.mean())
    combined_df_filled[combined_df_filled.columns[reverse_items]] = combined_df_filled[
    combined_df_filled.columns[reverse_items]].max() - combined_df_filled[combined_df_filled.
    columns[reverse_items]] + 1
    alpha_combined = pg.cronbach_alpha(data=combined_df_filled)
```

Listing 1: Calculate Cronbach's Alpha for combined dataset

Through the execution of the code on the data set, the Cronbach's alpha test yielded an alpha of ($\alpha = 0.8107$), indicating good internal consistency among the items. The 95% confidence intervals for Cronbach's Alpha were ranged ($95\%CI[0.656, 0.915]$)5, showing potential in general reliability, despite lower bound indicating a degree of uncertainty.

5.1.2 Test for Normality

For testing normality and assessing the normality of the data obtained, a Shapiro-Wilk test was conducted onto the test data of the two groups. The test was chosen due to its effectiveness regarding smaller sample sizes. The formula for the test is defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where:

- $x_{(i)}$ are the ordered sample values (i.e., the sample sorted in ascending order).
- $\bar{x}$ is the sample mean.
- a_i are constants generated from the means, variances, and covariance of the order statistics of a sample of size n from a normal distribution and are typically precomputed and tabulated for different sample sizes.

Based on the test the hypotheses tested were as follows:

- H_0 - The Data are drawn from a normally distributed population.
- H_1 - The data are not drawn from a normally distributed population.

Testing was done on both control ($n = 9$) and experimental ($n = 10$) data sets. The results from the test showed ($W = 0.970$, $p = .899$ for the control group. Similarly for the experimental group the test showed (W = 0.951, $p = .681$). A visualization can be seen at 13, in which the data is also shown as smoothed in order to account for the small sample size and illustrate the curvature of the distribution.

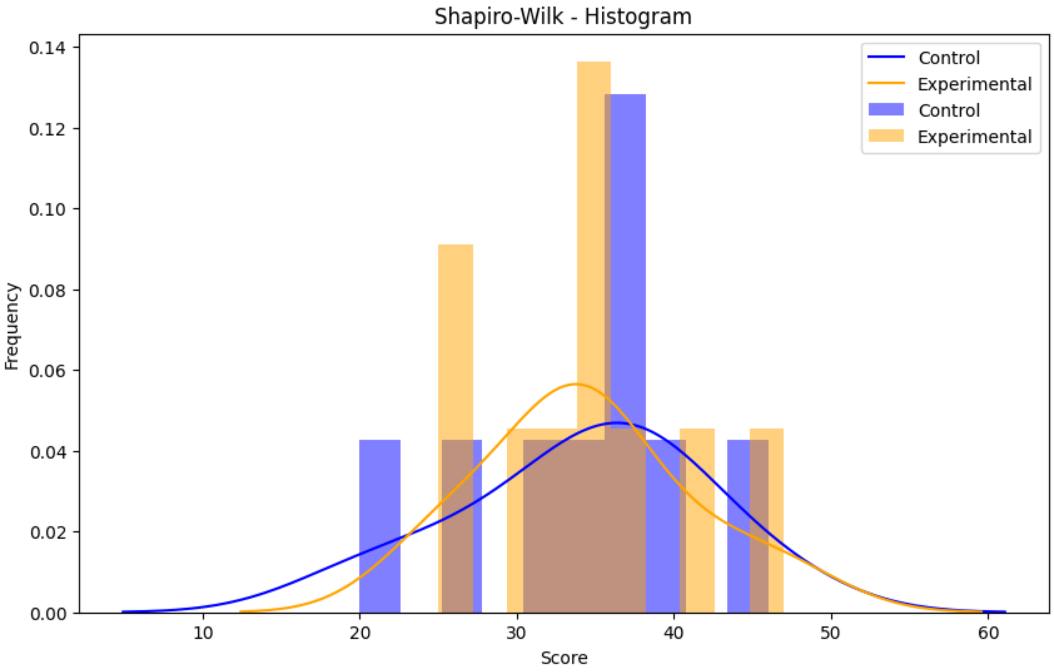


Figure 13: Graph of the distribution of the scores of the empathy questionnaire.

Given that both groups revealed a result that was greater than the significance level of 0.05, both groups both failed to reject the null hypothesis, pointing towards the probability that both groups were drawn from a normal distribution. As such, appropriate statistical analysis for the data should be employed.

5.1.3 Comparison of Means

Due to the data distribution assuming a normal distribution, the appropriate test for comparing the means between the two groups was planned to be an unpaired t-test. The test was conducted on both the control and experimental groups on their scores. The formula for the unpaired t-test is as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{X}_1$ and $\bar{X}_2$ are the sample means.
- s_1^2 and s_2^2 are the sample variances.
- n_1 and n_2 are the sample sizes.

Based on this, the hypotheses that were tested were as follows:

- $H_{0,1}$: AI NPCs do not significantly increase empathy in players compared to traditional NPCs.
- $H_{1,1}$: AI NPCs significantly increase empathy in players compared to traditional NPCs.

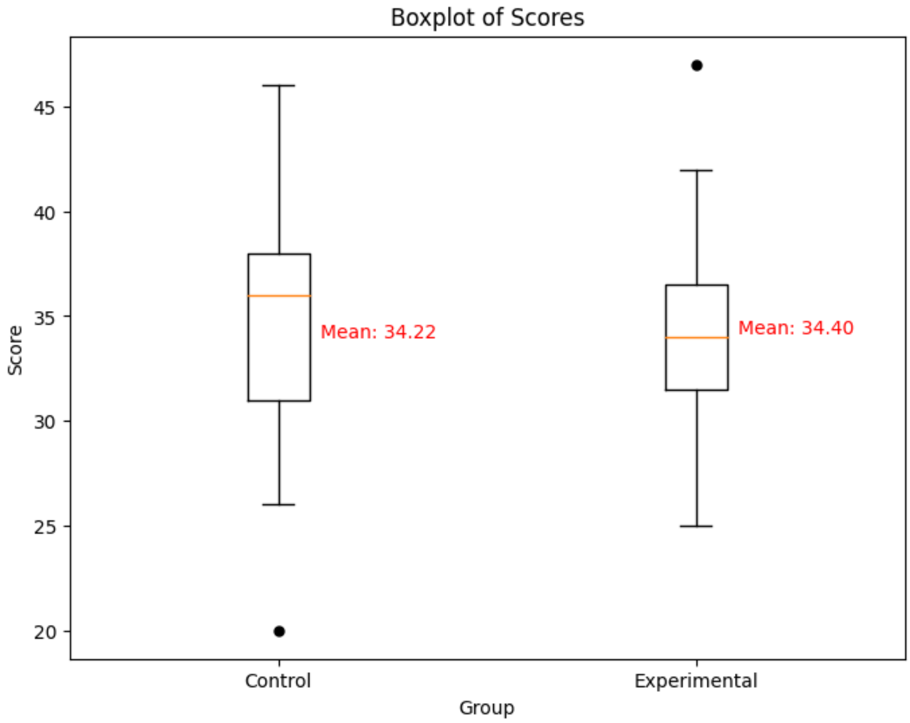


Figure 14: Boxplot showcasing the mean score of the two groups.

Shown in figure 14, the $(n = 10)$ participants who tested the experimental platform $(M = 34.4, SD = 6.3)$ compared to the $(n = 9)$ participants who tested the control platform $(M = 34.22, SD = 7.34)$ demonstrated a lack of change in scores. The test results of the t-test revealed that t(-0.05) the associated p-value of $(p = .957)$. Test revealed that $(p = .957)$ was greater than the significance level of 0.05, it fails to reject the null hypothesis. Therefore, there is no significant difference in the means of the control and experimental group.

5.2 Continuation Desire

After each participant concluded their play-though, they were asked to rate their desire to continue playing from a scale of 1 to 10. With 1 indicating a lack of desire to continue and 10 indicating a high desire to continue. The resulting dataset then underwent analysis to discover their results.

This served to answer the following hypothesis.

- $H_{0,2}$: AI NPCs do not significantly enhance player engagement compared to traditional NPCs.
- $H_{1,2}$: AI NPCs significantly enhance player engagement compared to traditional NPCs.

Table 1: Continuation Desire Rating Scoring

Table 2

Control	
Participant	Continuation Desire
P5	4
P6	3
P7	10
P8	5
P9	7
Avg	5.80

Table 3

Experimental	
Participant	Continuation Desire
P5	9
P6	10
P7	8
P8	7
P9	8
P10	10
Avg	8.667

The results of the continuation desire can be seen at an overview at table 1, in which each score obtained from the participants is stated as well as the calculated average. Like the data for empathy, the data was first tested for normality before assigning the appropriate statistical test.

5.2.1 Test for Normality

Once again Shapiro-Wilks was chosen as the appropriate test for assessing normality of the data.

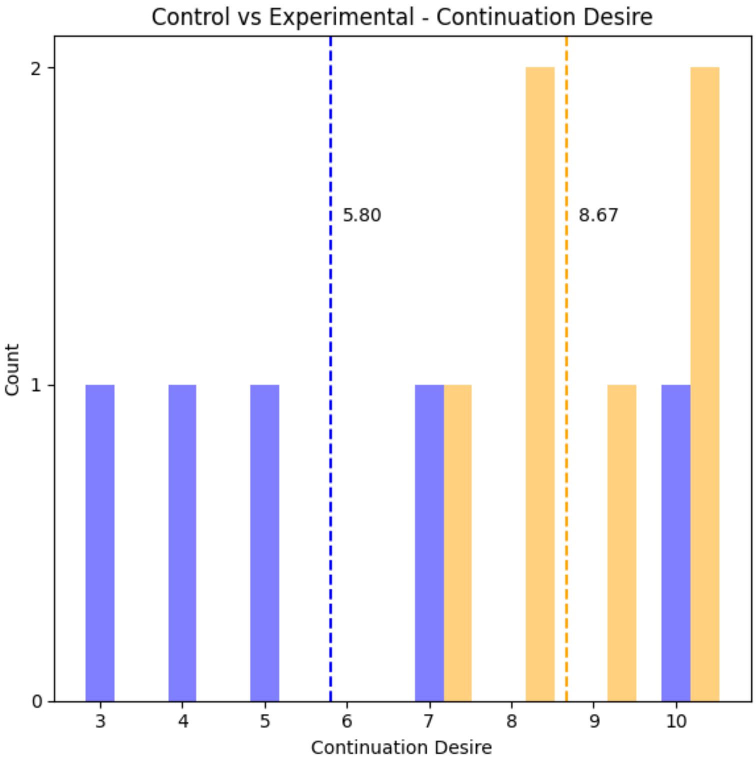


Figure 15: Bar-graph of the responses of the continuation desire between the two groups and their average.

5.2.2 Comparison of Means

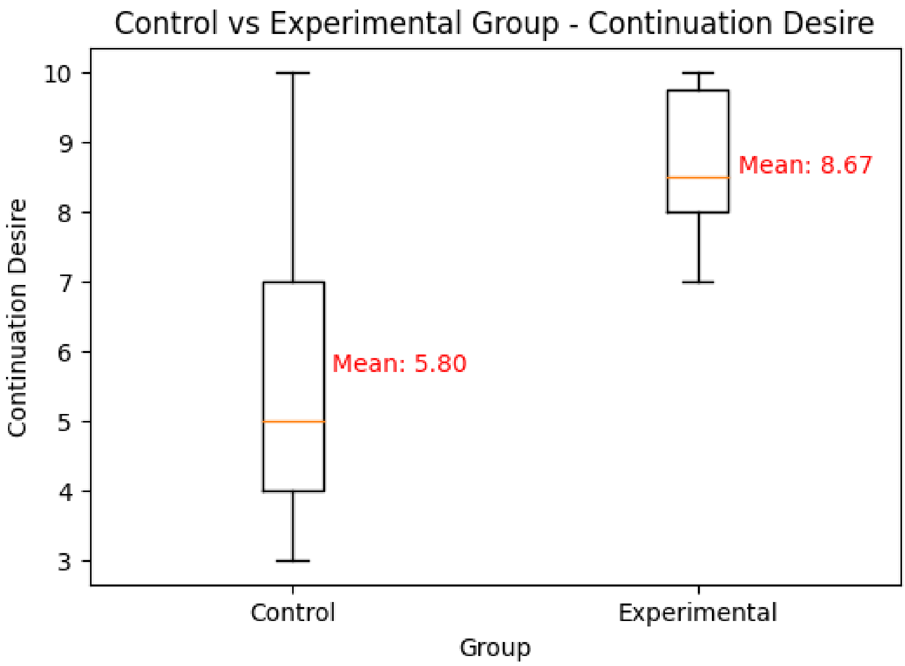


Figure 16: Box-plot showing the means for each group.

With the conformation of normality using the Shapiro-Wilks test, a T-test was conducted as the appropriate test for statistical difference. An unpaired t-test was performed to determine whether or not there was a significant difference in the continuation desire between the control group $(M = 5.8, SD = 2.8)$ and the experimental group $(M = 8.7, SD = 1.6)$. The test revealed a significant difference between the groups $(t(9) = -2.30, p = .047)$, indicating a notable difference in the continuation desire between the groups. Due to this significant difference, the null hypothesis can be rejected and accept the alternative hypothesis.

5.3 EMG

The EMG data from both groups were filtered to only include the last 200,000 samples and then the final 5000 of those samples were removed to avoid recording the participants reaction of the continuation desire question. The mean for each participant was plotted into a boxplot for the the individual sensors for each group:

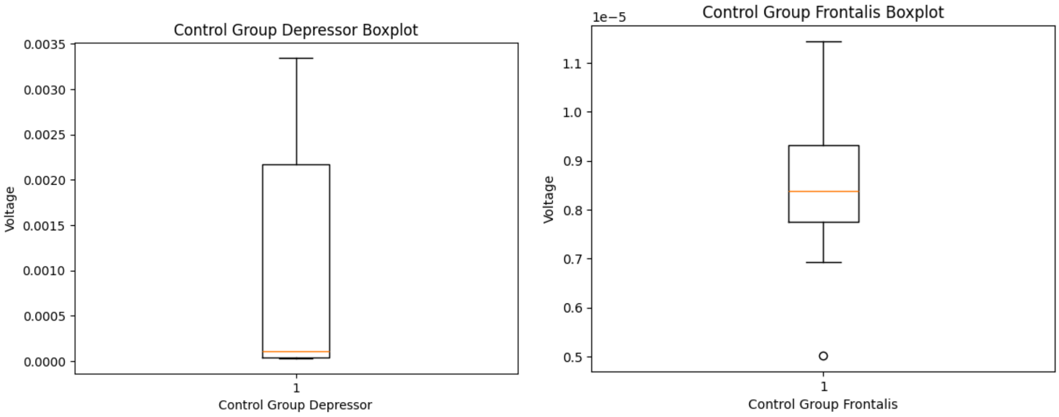


Figure 17: Boxplot for the controlgroups sensors.

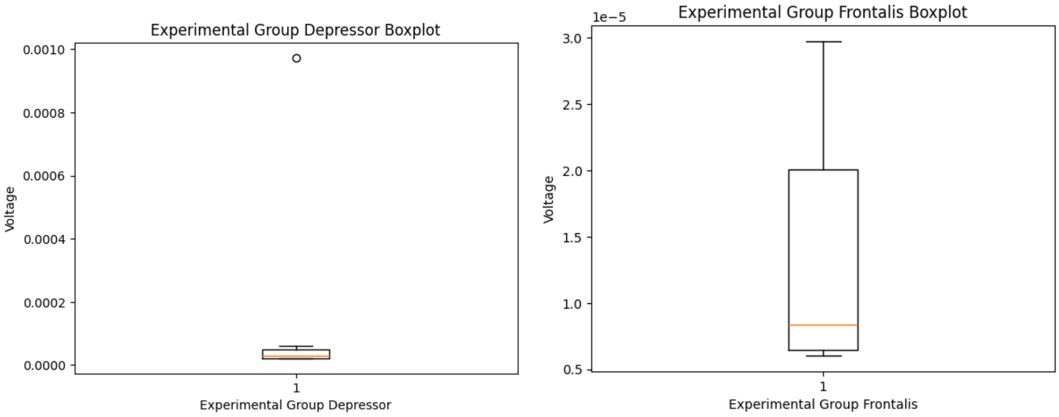


Figure 18: Boxplot for the experimentalgroups sensors.

On figure 17 and 18, the dots represents outliers within the data set. These data points were omitted from the set to get at better representation of the data. The Shapiro-Wilk test resulted in Control group Depressor Anguli Oris ($W = 0.720, p = .002$), Control group Frontalis ($W = 0.966, p = .866$), Experimental group Depressor Anguli Oris ($W = 0.833, p = .048$), and Experimental group Frontalis ($W = 0.773, p = .007$).

5.3.1 Comparison Test

Only one of the four groups showed to be normally distributed. Since the most of the groups are non-parametric, the Mann-Whitney U test was employed to see if there were a statistical difference between the experimental and control group for each sensor. The Mann-Whitney U test disregards the assumption of normality. The test with corresponding Mann-Whitney U statistic is defined as the smaller of:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_2$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_1$$

where:

- n_1 and n_2 are the sample sizes of the two groups.
- R_1 is the sum of the ranks of the observations in the first sample.
- R_2 is the sum of the ranks of the observations in the second sample.

U is then found be taking the lowest of

$$U = min(U_1, U_2)$$

The results of the Depressor Anguli Oris were (Mann–Whitney $U = 69, p = .011$) the Frontalis had the results: (Mann-Whitney $U = 42, p = .897$).

The mean was calculated for each of the categories "Depressor Anguli Oris" (M = $96.335 \times 10^{-5} v$), and "Occipitofrontalis"(M = $0.848 \times 10^{-5} v$). The experimental group had "Depressor Anguli Oris"(M = $12.795 \times 10^{-5} v$), and "Occipitofrontalis"(M = $1.295 \times 10^{-5} v$).

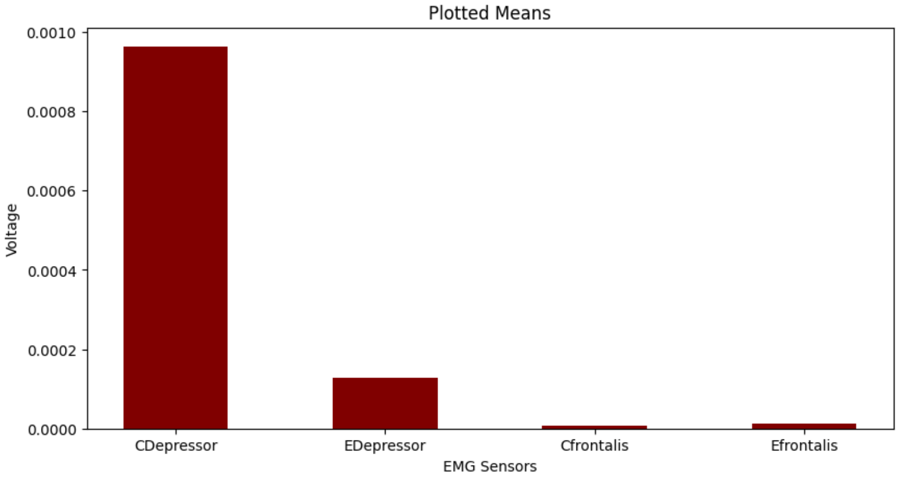


Figure 19: The means for each group Experimental(E) and Control(C).

The control group exhibited somewhat higher activity in the Depressor Anguli Oris muscle, which is commonly associated with expressions of sadness or disapproval. Opposite, the same group demonstrated slightly less activity in the Occipitofrontalis muscle.

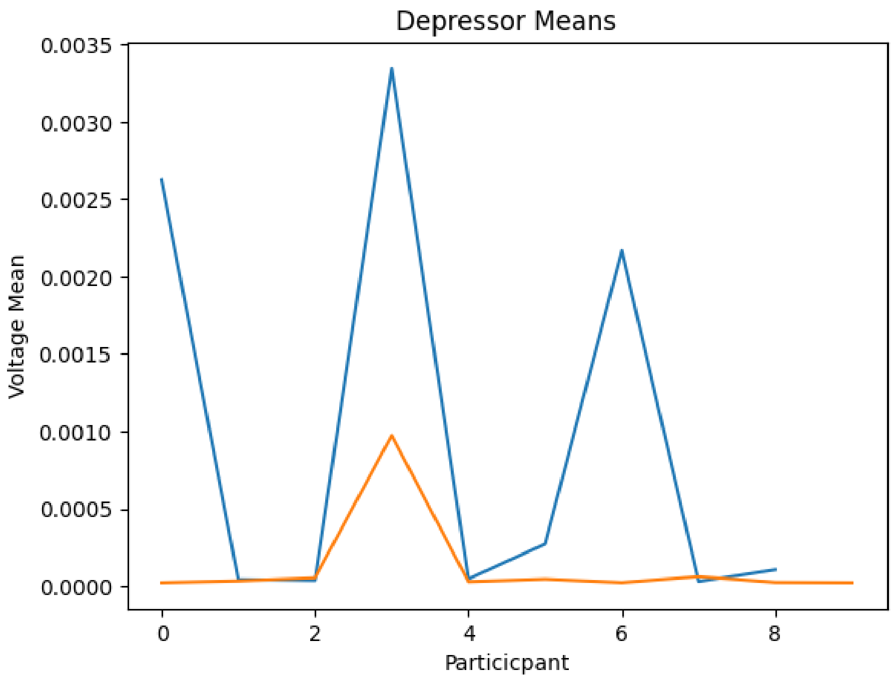


Figure 20: The avarage voltage for the Depressor on each participant, Control(blue) Experimental(Orange).

The measured voltage exhibited considerable variation between each participant in the control group for the Depressor Anguli Oris muscle. This variation highlights individual differences in muscle response, emphasizing the need to account for these differences for reliable results.

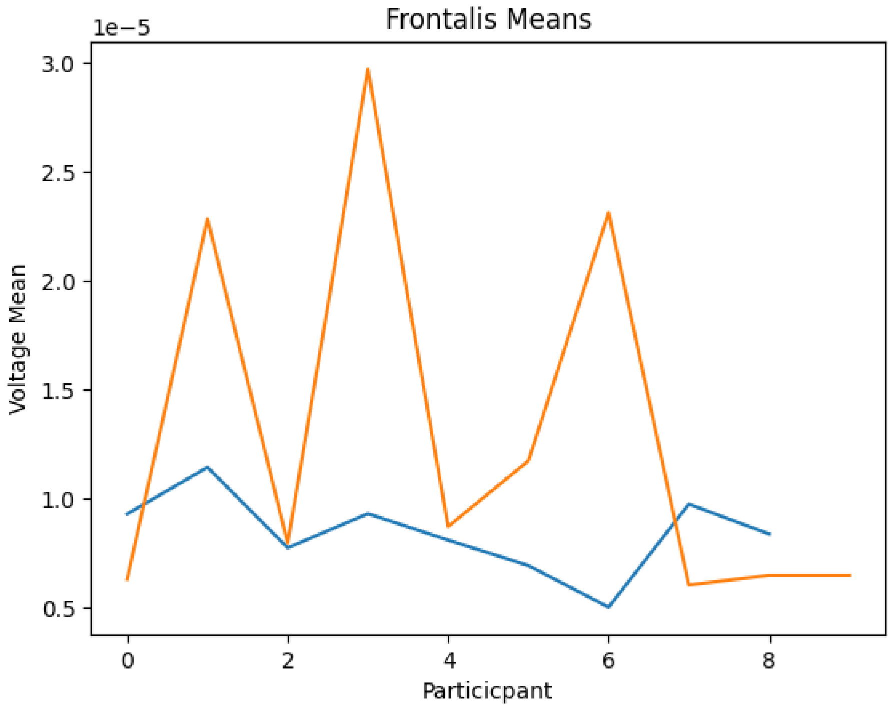


Figure 21: The average voltage for the Frontalis on each participant, control(blue) experimental(orange).

The experimental group had high variation for the measurement of the Occipitofrontalis.

6 Discussion

The purpose of this study was to examine differences in player engagement and empathy towards AI-driven versus traditional non-playable characters (NPCs) in video games.

The modifications made to the TEQ to fit for empathy towards a specific NPC, had its reliability tested through Chronbach's alpha. An alpha coefficient of 0.8107 was obtained from the analysis, which suggest that the items have a high degree of internal consistency. Conventionally, values above 0.8 indicate good reliability, while values above 0.7 are typically regarded as satisfactory. The calculated alpha value indicates that the modified version of the TEQ can still be considered to be a reliable instrument for measuring empathy. The high reliability is vital due to multiple reasons; It shows the internal consistency of the questionnaire that the items consistently measure the construct of empathy towards NPCs. Secondly, it adds to the overall credibility of the test and its results, and can prove as concept for future research benefiting from this or equal modifications to the TEQ. It is important to note that the overall response score of the modified version was lower compared to the original TEQ testing, which can be explained due to the removal of two items. It is furthermore possible that empathy in regards to NPCs simply is incapable of reaching same the level of empathy as living people, at its current state.

Testing and analyzing the responses shows that there are no significant difference in empathy between the groups. While this was the result of the experiment, it should not be held as a defining answer to whether or not users would be more or less prone to higher levels of empathy towards AI NPCs, as a lot of factors could play into these results. The experience itself was created with a specific intent of measuring a slight change of sadness in the user. However, the design was sub-optimal for the task, as designing for specific emotions can prove to be challenging. Furthermore, the gameplay was restricted to a five minute story and runtime, which severely limits the possibility of developing a proper relationship with *Bethy* compared to the time spend in most fully developed experiences and games. In regards to the chosen measurements, the emotional spectrum span a plethora of possible emotional states and combinations. So while testing for a single emotional state would suffice to reveal a partial change, much more thorough exploration of the emotional spectrum would have to be tested.

Through observations during testing of the platform, interesting participant behaviour was noted. While the participants in the experimental group had an infinite amount of content to experience, due to the open chat with *Bethy*, they would start to repeat the same questions and interactions. This might be a reflection of lack of guidance and purpose, as after the initial set up goals were completed participants relied solely om their own creativity for creating emergent stories with *Bethy*. This was further evident when some participants failed to activate the tasks and would wander aimlessly about in the café instead. Some participants expressed post testing that the scale of the experience was lacking in order to properly respond to the modified TEQ, as they did not feel they had experienced *Bethy* in the situations asked about through some of the items in the questionnaire. Feeling unsure of how to answer these items, participants could have been lead to skewed responses. One participant in particular showcased issue with the scale and capabilities of the experimental version, as the participant expressed their dis-satisfactory in *Bethy's* incapability of animation and serving of promised food and drinks.

Some participants were observed to purposefully attempt to break or otherwise work against the AI, none which led to output any harmful or ill-intended content, and only minor hallucinations occurred.

For the control version, some participants exhibited impatience with *Bethy*. By quickly skipping and clicking through the dialogue options, they either cut-off or completely missed what *Bethy* was saying. A behavior commonly seen when players are feeling less engaged or bored.

While there was shown to be no significant change between the two groups in regards to empathy, the most intriguing result came from the continuation desire. With a difference of means of 2.87 ($C = 5.8, E = 8.67$), and a statistical difference of ($p = .047$), there was a clear change between the two groups in favor of the experimental version. Rejecting the null hypothesis and showing promise in regards to AI NPCs improving the engagement of the user. These findings likewise serve as empirical data to support the findings from Inworld AI's research, that players prefer AI and their want to continue playing.

Some participants within the control group went through the experience by quickly clicking through and skipping most of the dialogue, interrupting *Bethy* and as an extension, shortening the experience drastically. These participants likewise sought out and completed the task given by *Bethy* at a fast pace, which resulted in them requiring to hold and wait until the finale. The common behaviour in these cases were running around the cafe, clicking at random stuff and expressing that they had no idea what to do.

A misunderstanding from the test-conductor led to the continuation desire being recorded wrongly initially. Instead of rating their desire to continue from a scale of one to ten, the they were simple asked whether or not the participants wished to continue, as a binary question. This meant that the continuation desire for the initial eight participants

could not be measured in correspondence to the remaining, as the binary data was incompatible with the quantitative ratings seen otherwise. Despite this, a noteworthy observation can be made, as three out of four in the control group said yes to continue. Conversely, only one of four participants from the experimental group expressed desire to continue. This showed a direct inverse expectation compared to results seen from the rated continuation desire, and could be due to a close four to six rating, that otherwise would be hard to measure.
Unfortunately, due to the low sample-size and the omitted responses of the initial eight participants, further testing would be preferred in order to provide a more consistent and reliant result. Despite this, the initial test is still able to show promise in regards to the future and potential of AI.

For the emotional response, EMG was employed in order to pick-up on bio-metric data that could potentially point towards the users subconscious emotional state. Results of the Mann-Whitney U test showed that there were a significant difference between the groups for the Depressor Anguli Oris with a value of ($p = .011$). The Frontalis showed no significant difference with a value of ($p = .897$) between the experimental and control group. The significant difference For the Depressor Anguli Oris could have been caused by noise since eight of the total participants had full or partial beard, in or close proximity to the Depressor Anguli Oris electrode. This affects the contact to the skin which is needed for more accurate readings, and the placement of the sensor on the muscle which can affect the results. As seen on 20 the measurements varied highly, and three of the participants had scores several times higher than the rest, due to the small sample size they were not considered outliers as they represent a third of the sample size. A similar trend can be seen for the Frontalis means on 21, here it is however the experimental group which appears to have a inconsistent voltage.

Participants 1 (index 0) and 7 (index 6), both had beards which might have caused the gap in readings, participant 4 (index 3) did not have a beard, creating some uncertainty on the reason for the extremely high value. The electrodes on the Frontalis were were placed on the forehead and therefore not affected by beard either, and the varying values were therefore more likely be caused by the electrodes not having the exact same placement and angle on the muscle affecting reading. The electrode to skin contact could also have affected the results, as lifting the eyebrows creates folds on the forehead making the surface uneven. Furthermore the level to which people emote with their face may also vary from person to person creating more signals. The participants went through the experience with the sound from speakers. Sound can have an effect on the readings, the participants did experience the same sound-clip, and only their distance from the screen could have affected the readings differently.
The overall testing revealed that EMG, while proven to be effective in measuring muscle movements and common use for bio-metric measurement, was too vague for measuring any meaningful responses in the facial region, and a different or additional approach to emotional response would be preferred. Alternatively further attempts with alternative EMG electrodes could provide better results.

Despite the findings of this study, it is important to consider a number of variables that could further affect the overall experience. The experience itself could have been different and by extension, potentially garnered a different result. Having a longer and more personalized adventure, could provide the user with a deeper connection to *Bethy*, and perhaps offer a changed view on empathy towards her. Furthermore, more complex system would allow for AI to showcase its possibilities with dynamic generation, offering even further unique and personalized experience. Due to its small scale, it would be difficult to predict the effect of using similar AI in larger productions seen in popular titles.

7 Conclusion

Multiple choice dialogue systems can often be tedious and cause players to even skip through the conversation, missing out on potential narrative elements. In role-playing games this effect can lead to constraining the player in a specific direction, and remove player agency. This paper set out to explore the effect using an AI-chatbot would have on the player compared to traditional NPC dialogue systems. Through literature and state of the art, a demonstration was shown how modern approach to narrative structure is seen in mainstream media and reflected for its incorporation with AI. Continuing the research of dynamic narrative experiences, such as *Façade*, the effect of implementing almost completely freedom of conversation proved to be interesting. Through the development of a narrative experience that implements an AI NPC, the effect of both player emotion, empathy, and engagement were measured in order to answer the research question:

How and to what extent does AI NPCs affect empathy and player engagement compared to traditional NPCs.

By modifying a validated questionnaire, the Toronto Empathy Questionnaire (TEQ), it would be possible to garner information in regards to the empathy towards NPCs, specifically the one created for the test of the study. It was shown that facial EMG is a a scientific field rooted back to the basic principles of Darwin's descriptions of emotions. For player engagement, the method of having the user rate their desire to continue, through continuation desire, would prove sustainable proof for player engagement.

These methods were applied to the testing of a platform developed for this study. The platform was created with two different versions. An experimental version, which allowed for free chat with the NPC, and a version that utilized the traditional dialogue system. The participants would have EMG sensors applied to their Depressor Anguli Oris and Frontalis and play through the experience, before responding to a questionnaire. The study had a total sample size of $N = 19$, with the experimental group consisting of $n = 10$ and the control group of $n = 9$. While previous study has proven its effectiveness, the results of the EMG would only show a significant difference for the Depressor Anguli Oris. These results however, could have been influenced due to general noisy data, that could interfere with the measurements. As such, it is difficult to conclude on the emotional effect the AI might have had on the participants. Nevertheless, the Depressor Anguli Oris indications show a slight favor towards the control group, while the Frontalis failed to yield a significant difference.

The Experimental and Control group showed no significant difference in terms of the modified TEQ, pointing towards no change in empathy when exposed to the experimental platform.

Continuation desire showed a significant difference between the two groups ($p = .047$), with the difference in means being 2.87 in favor of the experimental group. Showing that participants were more inclined to continue the experience while interacting with the AI compared to traditional dialogue system. However, due to a minor error, the sample size eligible for continuation desire was lower compared to the overall test. With the experimental having $n = 6$ and the control group having $n = 5$ participants.

Ultimately, the efforts and research of this study has achieved to prove potential in the improvement of engagement to a significant degree by implementing AI in a narrative driven experience, while bringing no change to the empathy of the player towards a given NPC.

7.1 Future Workflow

Due to the low sample size, there is a lack of concrete results. As such, further testing on a larger sample size would provide a more substantial level of clarity. Furthermore, changing the EMG electrodes to more accurate versions designed for facial EMG, as well as a more precise guide to placement would be beneficial. Facial EMG is also susceptible to interference from audio and environment noise. To circumvent this, headphones and an isolated testing area could prove beneficial for the EMG readings.

Implementing a machine-learning solution for facial emotion recognition could be a viable way to gather more data on the emotional state of the participant, while also being less intrusive compared to EMG.

If this would lead to indication of a larger effect size, further studies on individuals with social impairments, such as autism, could more accurately demonstrate a use-case for this technology. This would require further development of the prototype for allowing interaction with *Bethy* across different emotional stages. Researching the main challenges and obstacles would be essential for implementing an appropriate test procedure in order to test the level of ease of navigating intricate social cues and scenarios.

References

Ammanabrolu, Prithviraj et al. (2021). *How to Motivate Your Dragon: Teaching Goal-Driven Agents to Speak and Act in Fantasy Worlds*. arXiv: 2010.00685 [cs.CL].

Bakariya, B., A. Singh, and H. Singh (Apr. 2024). "Facial emotion recognition and music recommendation system using CNN-based deep learning techniques". In: *Evolving Systems* 15, pp. 641–658.

Bogost, Ian (2010). *Persuasive games: The expressive power of videogames*. mit Press.

Carvalhais, Tiago and Luís Magalhães (2018). "Recognition and Use of Emotions in Games". In: *2018 International Conference on Graphics and Interaction (ICGI)*, pp. 1–8. DOI: 10.1109/ITCGI.2018.8602898.

Cassell, J. et al. (1999). "Embodiment in conversational interfaces: Rea". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '99. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 520–527. ISBN: 0201485591. DOI: 10.1145/302979.303150. URL: https://doi.org/10.1145/302979.303150.

Cassell, Justine (2001). "Embodied conversational agents: representation and intelligence in user interfaces". In: *AI magazine* 22.4, pp. 67–67.

Chen, Vivian Hsueh-Hua et al. (2006). "Enjoyment or engagement? Role of social interaction in playing massively mulitplayer online role-playing games (MMORPGS)". In: *Entertainment Computing-ICEC 2006: 5th International Conference, Cambridge, UK, September 20-22, 2006. Proceedings 5*. Springer, pp. 262–267.

Dow, Steven (2007). "User engagement in physically embodied narrative experiences". In: *Proceedings of the 6th ACM SIGCHI Conference on Creativity & Cognition*, pp. 280–280.

Fraser, Jamie, Ioannis Papaioannou, and Oliver Lemon (2018). "Spoken Conversational AI in Video Games: Emotional Dialogue Management Increases User Engagement". In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. IVA '18. Sydney, NSW, Australia: Association for Computing Machinery, pp. 179–184. ISBN: 9781450360135. DOI: 10.1145/3267851.3267896. URL: https://doi.org/10.1145/3267851.3267896.

GDC, Game Developers Conference (June 2020). *Robocalypse Now: Using Deep Learning to Combat Cheating in Counter-Strike: Global Offensive*. Accessed: 29/02/2024. URL: https://www.youtube.com/watch?v=kTiP0zKF9bc.

Guzdial, Matthew and Mark Riedl (2016). "Game level generation from gameplay videos". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 12. 1, pp. 44–50.

Hafner, Manuela and Jeroen Jansz (2018). "The Players 'Experience of Immersion in Persuasive Games: The Players 'Experience of Immersion in Persuasive Games: A study of My Life as a Refugee and PeaceMaker". In: *International Journal of Serious Games* 5.4, pp. 63–79.

Hamari, Juho et al. (2016). "Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning". In: *Computers in human behavior* 54, pp. 170–179.

Inworld.Ai (Feb. 2023). *The Future of NPCs*. Accessed: 13-02-2024. URL: https://inworld.ai/whitepapers/future-of-npcs.

Ip, Barry (2011). "Narrative Structures in Computer and Video Games: Part 1: Context, Definitions, and Initial Findings". In: *Games and Culture* 6, pp. 103–134. URL: https://api.semanticscholar.org/CorpusID:146561104.

Ireland, David, Dana Bradford, and Geremy Farr-Wharton (2018). "Social Fringe Dwellers: Can chat-bots combat bullies to improve participation for children with autism?" In: *The Journal of Community Informatics* 14.

Ishii, Ryota et al. (2019). "A fighting game AI using highlight cues for generation of entertaining gameplay". In: *2019 IEEE Conference on Games (CoG)*. IEEE, pp. 1–6.

Jain, Vishal et al. (2020). "Algorithmic improvements for deep reinforcement learning applied to interactive fiction". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04, pp. 4328–4336.

Janelt, Tobias et al. (2023). "Analyzing the Factor Structure of the Toronto Empathy Questionnaire: Dimensionality, Reliability, Validity, Measurement Invariance and One-Year Stability of the German Version". In: *Journal of Personality Assessment*. DOI: https://doi.org10.1080/00223891.2023.2224873.

Jiang, Mingzhe et al. (Oct. 2015). "Facial Expression Recognition with sEMG Method". In: DOI: 10.1109/CIT/IUCC/DASC/PICOM.2015.148.

Kim, Mijin and Young Yim Doh (2017). "Computational Modeling of Players' Emotional Response Patterns to the Story Events of Video Games". In: *IEEE Transactions on Affective Computing* 8.2, pp. 216–227. DOI: 10.1109/TAFFC.2016.2519888.

Knickmeyer, Rachel Lee and Michael Mateas (2005). "Preliminary evaluation of the interactive drama facade". In: *CHI'05 extended abstracts on Human factors in computing systems*, pp. 1549–1552.

Kourmousi, Ntina et al. (June 2017). "The Toronto Empathy Questionnaire: Reliability and Validity in a Nationwide Sample of Greek Teachers". In: *Social Sciences* 6, p. 62. DOI: 10.3390/socsci6020062.

Krügel, Sebastian, Andreas Ostermaier, and Matthias Uhl (2023). "ChatGPT's inconsistent moral advice influences users' judgment". In: *Scientific Reports* 13.1, p. 4569.

Larian Studios (2023). *Baldur's Gate 3*. Various platforms. Video game.

Latitude-Inc. (2019). *AI Dungeon*. Accessed: 28-02-2024. URL: https://play.aidungeon.com.

Lessard, Jonathan (2016). "Designing natural-language game conversations". In: *Proc. DiGRA-FDG* 16.

LokeshNaik, S.K. et al. (2023). "Real Time Facial Emotion Recognition using Deep Learning and CNN". In: *2023 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–5. DOI: 10.1109/ICCCI56745.2023.10128259.

Lyons, Elizabeth J et al. (2014). "Engagement, enjoyment, and energy expenditure during active video game play." In: *Health Psychology* 33.2, p. 174.

Mateas, Michael and Andrew Stern (2003). "Façade: An experiment in building a fully-realized interactive drama". In: *Game developers conference*. Vol. 2. Citeseer, pp. 4–8.

Matsumoto, David et al. (2010). "Facial expressions of emotion". In: vol. 3. Handbook of Emotions. Chap. 13, pp. 211–234.

Mori, Yusuke and Youichiro Miyake (2022). "Ethical Issues in Automatic Dialogue Generation for Non-Player Characters in Digital Games". In: *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 5132–5139.

Naul, Emily and Min Liu (2020). "Why story matters: A review of narrative in serious games". In: *Journal of Educational Computing Research* 58.3, pp. 687–707.

Neumann, David. L. et al. (2014). *Measures of Empathy: Self-Report, Behavioral, and Neuroscientific Approaches*. Elsevier Science & Technology, pp. 257–284.

OpenAI (Mar. 2021). *Dota2 with Large Scale Deep Reinforcement Learning*. URL: https://arxiv.org/pdf/1912.06680.pdf.

Qin, Hua, Pei-Luen Patrick Rau, and Gavriel Salvendy (2009). "Measuring player immersion in the computer game narrative". In: *Intl. Journal of Human–Computer Interaction* 25.2, pp. 107–133.

Quantic Dream (2018). *Detroit: Become Human*. PlayStation 4. Video game.

Radford, Alec et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9.

Rashkin, Hannah et al. (2020). *PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking*. arXiv: 2004.14967 [cs.CL].

Ryan, Marie-Laure (1991). *Possible worlds, artificial intelligence, and narrative theory*. Indiana University Press.

— (2015). *Narrative as virtual reality 2: Revisiting immersion and interactivity in literature and electronic media*. JHU press.

Santiago III, Jose Ma et al. (2023). "Rolling the dice: Imagining generative ai as a dungeons & dragons storytelling companion". In: *arXiv preprint arXiv:2304.01860*.

Schoenau-Fog, Henrik (2011a). "Hooked! – Evaluating Engagement as Continuation Desire in Interactive Narratives". In: *Interactive Storytelling*. Ed. by Mei Si et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 219–230. ISBN: 978-3-642-25289-1.

— (Jan. 2011b). "The Player Engagement Process - An Exploration of Continuation Desire in Digital Games". In: *DiGRA &#3911 - Proceedings of the 2011 DiGRA International Conference: Think Design Play*. DiGRA/Utrecht School of the Arts. URL: http://www.digra.org/wp-content/uploads/digital-library/11307.06025.pdf.

Schoenau-Fog, Henrik et al. (2013). *Narrative engagement in games–a continuation desire perspective*. Society for the Advancement of the Science of Digital Games.

Sifa, Rafet et al. (2013). "Behavior evolution in Tomb Raider Underworld". In: *2013 IEEE Conference on Computational Inteligence in Games (CIG)*, pp. 1–8. DOI: 10.1109/CIG.2013.6633637.

Sobieszek, Adam and Tadeusz Price (2022). "Playing games with AIs: the limits of GPT-3 and similar large language models". In: *Minds and Machines* 32.2, pp. 341–364.

Spreng, R. N. et al. (2009). "The Toronto Empathy Questionnaire: scale development and initial validation of a factor-analytic solution to multiple empathy measures". In: *Journal of personality assessment*. DOI: https://doi.org/10.1080/00223890802484381.

Strong, Christina and Michael Mateas (2008). "Talking with NPCs: Towards dynamic generation of discourse structures". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 4. 1, pp. 114–119.

Takahashi, Tomomi, Kazuaki Tanaka, and Natsuki Oka (2018). "Adaptive Mixed-Initiative Dialog Motivates a Game Player to Talk with an NPC". In: *Proceedings of the 6th International Conference on Human-Agent Interaction*. HAI '18. Southampton, United Kingdom: Association for Computing Machinery, pp. 153–160. ISBN: 9781450359535. DOI: 10.1145/3284432.3284436. URL: https://doi.org/10.1145/3284432.3284436.

Taveekitworachai, Pittawat et al. (2023). "Breaking bad: Unraveling influences and risks of user inputs to chatgpt for game story generation". In: *International Conference on Interactive Digital Storytelling*. Springer, pp. 285–296.

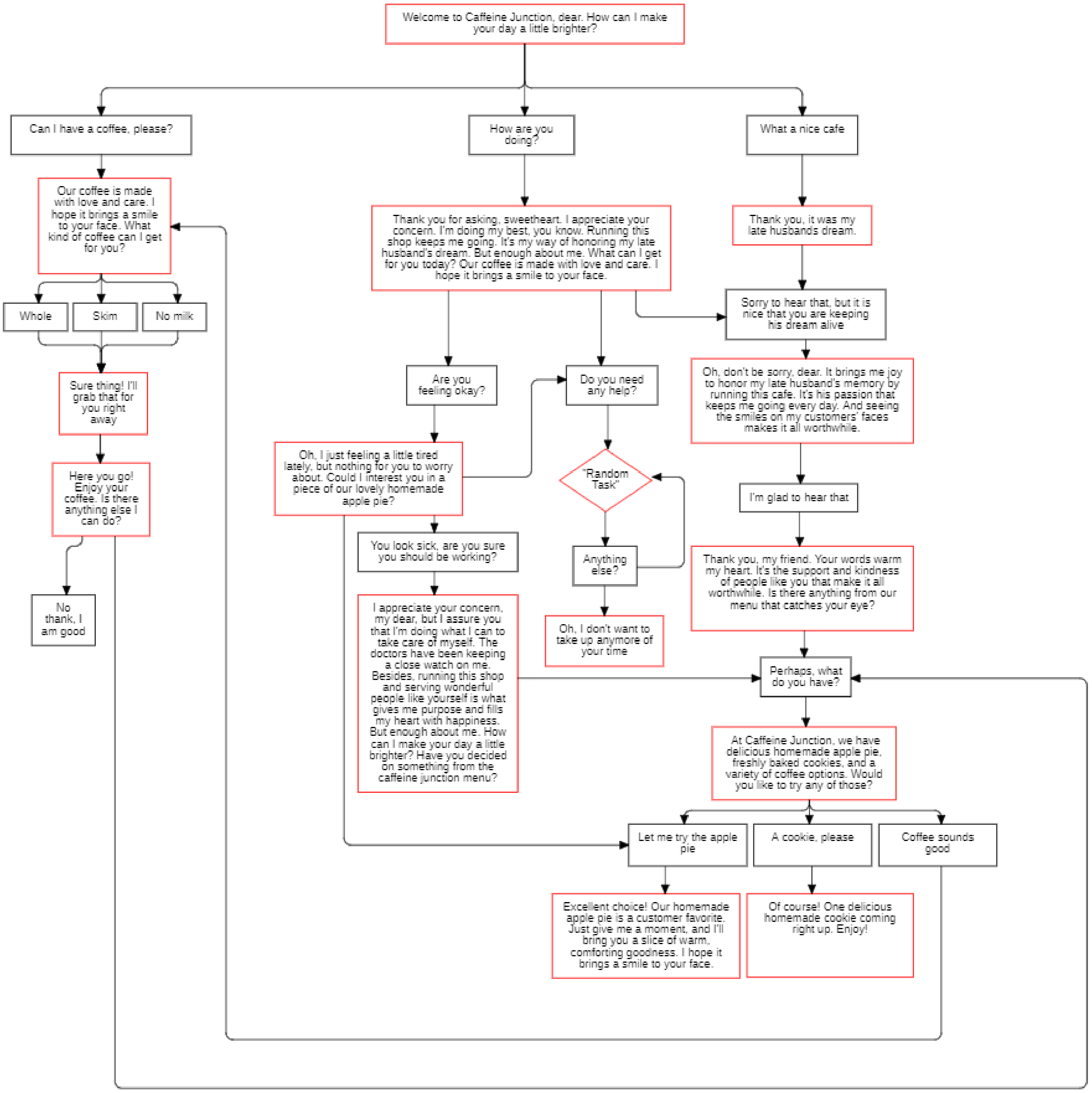
Totan, Tank, Doğan Tayfun, and Fatma Sapmaz (Dec. 2012). "The Toronto Empathy Questionnaire: Evaluation of Psychometric Properties among Turkish University Students". In: *Egitim Arastirmalari - Eurasian Journal of Educational Research* 12, pp. 179–198.

Willerton, Chris (2000). "Structure problems in hypertext mysteries". In: *Proceedings of the eleventh ACM on Hypertext and hypermedia*, pp. 234–235.

Xu, Richard Huan et al. (2020). "Validation of the Toronto Empathy Questionnaire (TEQ) Among Medical Students in China: Analyses Using Three Psychometric Methods". In: *Sec. Quantitative Psychology and Measurement*. DOI: https://doi.org/10.3389/fpsyg.2020.00810.

Yang, Songfan and Bir Bhanu (2012). "Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4, pp. 980–992. DOI: 10.1109/TSMCB.2012.2192269.

A Appendix A


Welcome to Caffeine Junction, dear. How can I make your day a little brighter?
Can I have a coffee, please?
How are you doing?
What a nice cafe
Our coffee is made with love and care. I hope it brings a smile to your face. What kind of coffee can I get for you?
Thank you for asking, sweetheart. I appreciate your concern. I'm doing my best, you know. Running this shop keeps me going. It's my way of honoring my late husband's dream. But enough about me. What can I get for you today? Our coffee is made with love and care. I hope it brings a smile to your face.
Thank you, it was my late husbands dream.
Whole
Skim
No milk
Sorry to hear that, but it is nice that you are keeping his dream alive
Are you feeling okay?
Do you need any help?
Oh, don't be sorry, dear. It brings me joy to honor my late husband's memory by running this cafe. It's his passion that keeps me going every day. And seeing the smiles on my customers' faces makes it all worthwhile.
Sure thing! I'll grab that for you right away
Oh, I just feeling a little tired lately, but nothing for you to worry about. Could I interest you in a piece of our lovely homemade apple pie?
"Random Task"
I'm glad to hear that
Here you go! Enjoy your coffee. Is there anything else I can do?
You look sick, are you sure you should be working?
Anything else?
Thank you, my friend. Your words warm my heart. It's the support and kindness of people like you that make it all worthwhile. Is there anything from our menu that catches your eye?
No thank, I am good
I appreciate your concern, my dear, but I assure you that I'm doing what I can to take care of myself. The doctors have been keeping a close watch on me. Besides, running this shop and serving wonderful people like yourself is what gives me purpose and fills my heart with happiness. But enough about me. How can I make your day a little brighter? Have you decided on something from the caffeine junction menu?
Oh, I don't want to take up anymore of your time
Perhaps, what do you have?
At Caffeine Junction, we have delicious homemade apple pie, freshly baked cookies, and a variety of coffee options. Would you like to try any of those?
Let me try the apple pie
A cookie, please
Coffee sounds good
Excellent choice! Our homemade apple pie is a customer favorite. Just give me a moment, and I'll bring you a slice of warm, comforting goodness. I hope it brings a smile to your face.
Of course! One delicious homemade cookie coming right up. Enjoy!

B Appendix B

Many techniques for reading emotions can be intrusive which may affect the results, therefore an unintrusive method could help the participants get the intended experience during play. One way could be through image recognition where the participant can be recorded during play(Yang & Bhanu, 2012), facial expressions in video have also been used in real time for a simple game, where the game-state would change based on the participants expression, it is however still hard to pinpoint the exact emotion as emotional states can be complex(Carvalhais & Magalhães, 2018). The facial expressions may also vary from person to person as some might have a wider smile while other have a more discrete smile, which could cause trouble for algorithms trying to classify the given emotion objectively(Carvalhais & Magalhães, 2018). Other research shows that a successful game may trigger more emotional responses more often, as research shows that the successful games would trigger emotional responses up to 3.3 times more often than the unsuccessful ones(Kim & Doh, 2017).

Some research limit the number of number of emotions to reduce the amount of classes that might have to be categorised. Different papers and techniques does however use some variations of the same emotions (happiness, surprised, sadness, fear, anger, disgust, and neutrality)(Carvalhais & Magalhães, 2018) (LokeshNaik et al., 2023),(Bakariya, Singh, & Singh, 2024) uses the same emotions but have chosen to leave out neutrality,(Anger, Contempt, Disgust, Fear, Happy, Sadness, and Surprise)(Yang & Bhanu, 2012). Most of these examples ended with an accuracy of between 70 to 80 percent, which is a little on the low side as one in four facial expressions could be categorised wrongly, having a classifier that separates the expressions into positive and negative emotions, before classifying the exact emotion can improve accuracy, As fear and happiness can look similar(Carvalhais & Magalhães, 2018).

C Appendix C

Script:
Hello and welcome, Thank you for participating in our test. For the purpose of equal treatment we will conduct the test in English, but feel free to respond in Danish if you prefer. Before we start, we'll need you to fill out some basic information on the computer in front of you and ask for your consent. After you've completed the form, we'll give you some wet wipes to clean your face in the areas where we'll place the EMG sensors. This is to ensure we get a good signal. Please use the wipes to gently clean "AREAS".

(After they have wiped the area, attach sensors) After placing sensors: In a second i will start the game, I'll just give you a quick guide on how to play and navigate

1. Use 'WASD' to move around.
2. Move the mouse to look around.
3. Press 'The left mouse button' to interact with things, you should get a prompt if you can interact with something.

You'll have about 5 minutes to play the game, but I will interrupt you when you are finished. During that time, please try to sit still and avoid moving around too much, as it can affect the sensors. Just try and focus on playing the game. You are free to ask questions should you require assistance, but otherwise, do refrain from talking too much. Any questions?

[Starting EMG recording]
[Start the Game]
(Take notes during testings)
(After End of Game)

[Close the game and stop recording]
That's it, I hope you had a good time, I will now remove the sensors.
[Remove sensors]
If you had to rate on a scale from 1 to 10, how interested would you have been in continuing to play this game or interact with Bethy? 1 means you're not interested at all, and 10 means you would love to keep playing.
[Note answer down, including any comments made].
You're all done. You can now continue with the final questionnaire. **After questionnaire:**

Thank you for participating.

D Appendix D

Appendix D is a .zip folder attached separately titled "Appendix D".