Bachelor thesis

Artificial Intelligence

Radboud University

The effect of humour on the user experience of a chatbot within a gaming environment

Author:
Maud van de Lockant
S1009396

First supervisor:
M. A. de Sá Siqueira
PhD student - Behavioural Science Institute
marianna.desasiqueira@ru.nl

Second supervisor:
prof. dr. T. Bosse
Behavioural Science Institute
t.bosse@ru.nl

June 18, 2021

Abstract

Chatbots are becoming increasingly popular within the world of Artificial Intelligence and natural language. They can have a multitude of purposes, such as customer service, social and emotional support, information, entertainment, and redirection. The problems faced by chatbots, like being irritating, repetitive or intimidating, resemble obstacles where humans use humour to overcome. Additionally, humans interact similarly with computers as with other humans. During the experiment, a small-talk text-based chatbot was developed that can play the game Connect Four. A chatbot with neutral, polite dialogue was compared to a chatbot with added forms, puns and humorous remarks, of humour. The user experience was measured by using the Technology Acceptance Model. This way, this thesis tries to provide insight into the role of humour on the user experience within a gaming environment. The results suggest that there is no significant effect of humorous responses on the user experience when using the chatbot. Nevertheless, there is a significant correlation between Perceived Ease of Use and Perceived Usefulness, Perceived Usefulness & Perceived Ease of Use and Attitude towards Using, Perceived Usefulness & Attitude towards Using and Behavioural Intention to Use of the Technology Acceptance Model.

Contents

1 Introduction **3**

2 Background **5**
2.1 Chatbot 5
2.2 Humour 5
2.2.1 Functionalities of humour 5
2.2.2 Potential pitfalls for humour 6
2.2.3 Combination natural language interfaces and humour 7
2.2.4 Relation between humour and user experience 8
2.3 Connect four 8
2.4 Technology acceptance model 9
2.4.1 Perceived Usefulness 10
2.4.2 Perceived Ease of Use 10
2.4.3 Attitude Towards Using 11
2.4.4 Behavioural Intention to Use 11
2.5 Research model and hypotheses 11

3 Method **13**
3.1 Resources 13
3.1.1 Website Design 13
3.1.2 Chatbot 15
3.1.3 Connect Four 19
3.1.4 Questionnaire 19
3.2 Experiment design 20

4 Results **22**
4.1 Quantitative results 22
4.1.1 Outlier removal 22
4.1.2 Participants 23
4.1.3 Visualisation 24
4.1.4 ANOVA 27
4.1.5 Pearson's correlation coefficient 27
4.2 Qualitative results 30

5 Discussion **32**

6 Conclusions **35**

7 Acknowledgements **36**

Bibliography **37**

A Informed consent form **42**

B Task Instructions **44**
B.1 Website Tutorial Text 45

C Questionnaire **46**
C.1 General questions 46
C.2 TAM-related questions 48
C.3 Open questions 50
C.4 Debriefing Text 50

D Connect Four Implementation **51**

Chapter 1

Introduction

How many people are there these days who haven't used either Apple's Siri, Microsoft's Cortona, Amazon's Alexa or Google's Assistant? These virtual agents are becoming increasingly popular within the world of Artificial Intelligence and natural language. Although the first conceptualized problem as an "imitation game" of a virtual agents had already been made in the 1950s by Alan Turing (Turing, 1950), chatbots only recently started to make it more and more into our day-to-day life. These agents also go by other names such as chatbots, conversational agents, or digital assistants. For the rest of this thesis, I will simply refer to these agents as chatbots.

The ultimate goal of a chatbot is to mimic human-to-human interaction as closely as possible. This can also be seen by the set-up of the Turing test, namely the ability of the chatbot to be mistakenly judged as a real human being by a user (Turing, 1950).

Chatbots consist of user interfaces, in which interaction with the digital systems happens through either text or speech in natural language (Følstad and Brandtzæg, 2017). When in a human-to-human interaction using natural language most people use some form of humour every day (Binsted, 1995). The problems faced by natural language interfaces, such as being irritating, repetitive or intimidating, resemble obstacles where humans use humour to overcome (Binsted, 1995). Additionally, users respond to computers in the same way that they respond to other people (Reeves and Nass, 1996). This makes humour a significant subject to research when trying to achieve the goal of mimicking a human-to-human interaction.

Accordingly, this thesis proposes the following research question to find a solution for this problem. *What is the effect of humour on the user experience of a chatbot within a gaming environment?*

Previously research into computers as social actors by Nass et al. (1996), Nass et al. (1999), Nass and Moon (2000) and Reeves and Nass (1996)

shows that people tend to like computers more when they used humour. Another study by Morkes et al. (1998) demonstrates that computer agents that use humour are rated as more likeable, competent and cooperative than those who do not use humour. An indication of the importance of humour is the chatbot created by Microsoft named Xiaoice. She has about 20 million registered users. They say to be drawn to her for her sense of humour and listening skills (D'onfro, 2015).

To evaluate the research question there were two chatbots developed, one neutral and polite and the other enhanced with humorous responses. Both these chatbots co-occurring play the game Connect Four. The user experience will be measured via a questionnaire based on the Technology Acceptance Model. The research model, as well as the hypotheses, will also be based on this model. The data will be tested against each other with the statistical methods ANOVA and Pearson's correlation coefficient.

Chapter 2

Background

2.1 Chatbot

The first chatbot created, after the initial conceptualized problem as an "imitation game" (now called the Turing Test) (Turing, 1950), was ELIZA in 1966 (Weizenbaum, 1966). This computer program was developed to mimic the responses of a psychotherapist in a therapy session (Weizenbaum, 1966). These days chatbots can have a multitude of purposes, such as customer service, social and emotional support, information, entertainment, and redirection to other people or machines (Brandtzaeg and Følstad, 2017). In particular, Xu et al. (2017) shows that chatbots can be seen as a promising alternative to traditional customer service. It may be more natural for customers to use natural language than use a mobile app for questions regarding products and services (Brandtzaeg and Følstad, 2017).

Based on the original Turing test, the Loebner prize was created in 1991. The prize is handed out to whoever creates the chatbot which is most wrongly identified as human (Bradeško and Mladenić, 2012). The best-known winners are Cleverbot, A.L.I.C.E., and the more recent Mitsuku (Brandtzaeg and Følstad, 2017). There is also some debate around the real contribution to chatbots of the Loebner Prize. This mostly relates to the idea that a chatbot can pretend to be human without real intelligence. Examples of this are forced spelling mistakes (Bradeško and Mladenić, 2012).

2.2 Humour

2.2.1 Functionalities of humour

The pervasiveness of humour in human social relationships (Martin and Ford, 2018) is not without reason. Graham et al. (1992) name a number of important functions of humour which arose from previous research. Such functions include (not limited to):

1. Allow others insight into another's state of mind (Civikly, 1983) (Linstead, 1985).

2. To decrease another's aggressive behaviour (Baron and Ball, 1974) (Civikly, 1983) (Civikly, 1989) (Whitaker, 1975).

3. To disclose difficult information (Civikly, 1983) (Civikly, 1989) (Smith et al., 1971).

4. To help others relax and feel comfortable (Civikly, 1983) (Civikly, 1989) (Landy and Mettee, 1969) (Smith and Powell, 1988).

5. To entertain others (Civikly, 1983) (Civikly, 1989) (Stocking and Zillmann, 1976) (Weaver et al., 1988) (Zillmann and Bryant, 1983).

6. To minimize anxiety (Bricker, 1980) (Civikly, 1983) (Civikly, 1989) (Smith and Powell, 1988).

7. To reduce boredom (Civikly, 1983) (Civikly, 1989) (Roy, 1959).

8. To facilitate relationship patterns (Sykes, 1966).

Humour can be spontaneous (such as making an accidental spelling error) or be deliberate (such as predefined jokes). Humour happens within conversations with friends but also in more formal environments like teaching environments and the workplace (Nijholt, 2007). Within these formal environments, a lot of positive effects were perceived (Nijholt, 2007). Even that much that big companies like IBN and AT&T have hired humour professionals to improve features like teamwork and motivation (Gibson, 1994).

2.2.2 Potential pitfalls for humour

While the use of humour is universal, the sense of humour is highly shaped by cultural environment (Ruch, 2007). Jokes that would provoke laughter within an Indian environment would hardly draw a smile from a Dutch person (Driessen, 2015). This holds the same for aggressive humour. Americans seem to prefer this more than their Belgians, Hong-Kongese (Castell and Goldstein, 1977), Senegalese and Japanese counterparts do (Goldstein et al., 1976). Even within groups speaking the same language, there could be cultural differences. An example of this is that native Hungarians show more appreciation for jokes featuring ethnic stereotypes compared to bilingual English-Hungarians (Erdodi and Lajiness-O'Neill, 2012) (Martin and Sullivan, 2013). Ptaszynski et al. (2010) states that puns are for example one of the main humour genres in Japanese culture, but puns are language-specific and difficult to translate. As such, it is important to take your users' social norms, culture-specific conventions, and language into account (Chaves and Gerosa, 2019). Also to share a common background with the

audience and master language subtlety when designing forms of humour which are appropriate (Reimann et al., 2010).

Personal and cultural differences explain why many jokes or ironic remarks often go unremarked, misunderstood, or perceived as offensive (Reimann et al., 2010). This raises additional questions to look out for, for example: How many jokes are deemed too many? In which situations is humour abjectly inappropriate? Are unsuccessful jokes a problem (Morkes et al., 1998)?

One of the answers is stated by Adams (1995) which says: humour can ease interaction but inappropriate humour can be unbearably irritating. Ethnical, racial, and sexist jokes have offensive connotations. Additionally, intellectual jokes might not be always understood by all people, while self-deprecating jokes can affect the image of oneself (Niculescu et al., 2013). Also, dark, vulgar, or toilet humour should be avoided, since it suggests an interlocutor with rather low class attributes (Nass and Brave, 2005). So when implementing computational humour into a chatbot to make it more friendly it must be in the right context and fitted for the user (Binsted, 1995). This especially holds when it is a frequent user of the system (Binsted, 1995).

2.2.3 Combination natural language interfaces and humour

Despite the positive effects and functionalities of humour in personal and work situations, the Human-Computer Interaction (HCI) field still keeps a rather negative perspective towards the use of humour in interfaces (Niculescu et al., 2013). Since the basics of HCI policies are maximizing efficiency and minimizing time, humour would be useless and could be even called distracting (Niculescu et al., 2013). Fortunately, within the artificial intelligence and natural language processing world, humour is a lot better established (Niculescu et al., 2013). They see the use of humour and the improvements that could be made towards humanizing chatbots. Binsted (1995) agrees upon this view and states that the important functions of humour may reduce the difficulties that are presented by natural language interfaces such as chatbots. Also when a chatbot needs to recover from a made error, humour can be useful. A study by Niculescu and Banchs (2015) shows that using humour in a time of error may stimulate the user to rephrase his entry.

Due to the pitfalls introduced by humour, multiple researchers Ptaszynski et al. (2010) Gibson (1994) Morkes et al. (1998) Medhi Thies et al. (2017) Augello et al. (2011) Dybala et al. (2008) Tseng et al. (2020) Khooshabeh et al. (2011) discuss chatbots who use puns and jokes instead of complex humour to create funny conversations. Niculescu et al. (2013) Augello et al. (2008) expanded upon these kinds of humour with ironic answers and riddles. Blinov et al. (2017) discusses another approach to the complexity namely generation of funny Tweets.

2.2.4 Relation between humour and user experience

A multitude of researchers looked into the application of humour in chatbots. Firstly, Huan and Szafir (2001) findings revealed that adding humour enhanced the virtual instructor's likeliness significantly. This was even the case regardless of whether the instructor was a human or a robot. Secondly, Niculescu et al. (2013) found that humour in a conversation with a social robot receptionist increased users' liking for the robot's speaking style and personality, as well as task enjoyment overall. Nextly, users interacting with a conversational agent designed by Dybala et al. (2009) viewed the humorous chatbot as being more funny and likeable, as well as more human-like. When they compared the two agents, humorous versus non, they discovered that the humorous agent received higher ratings. Last but not least, Babu et al. (2006) discovered that by utilizing jokes, a humorous virtual receptionist can encourage users to participate more in social dialogue.

Also, as described in the introduction, Morkes et al. (1998) discovered that participants who received humorous remarks from a computer viewed the system as much more cooperative, likeable, and competent. Additionally, those who received the humorous comments also smiled and laughed more during the trial, demonstrating greater social behaviour.

Another relating observation made by Khooshabeh et al. (2011) is that when participants perceive a chatbot to be humorous then the chatbot is more effective at socially influencing users.

All of this previous research suggests that adopting humour in a chatbot's conversation improves the chatbot's likeability and may contribute to a better user experience (Niculescu and Banchs, 2019).

2.3 Connect four

Connect Four, also known as four in a row, is a board game played by two players. The players take turns dropping coloured discs into the board. The player which has four adjacent disks (horizontally, vertically, or diagonally) first wins the game. The game is originally played in a seven-by-six grid but some variants were later made.

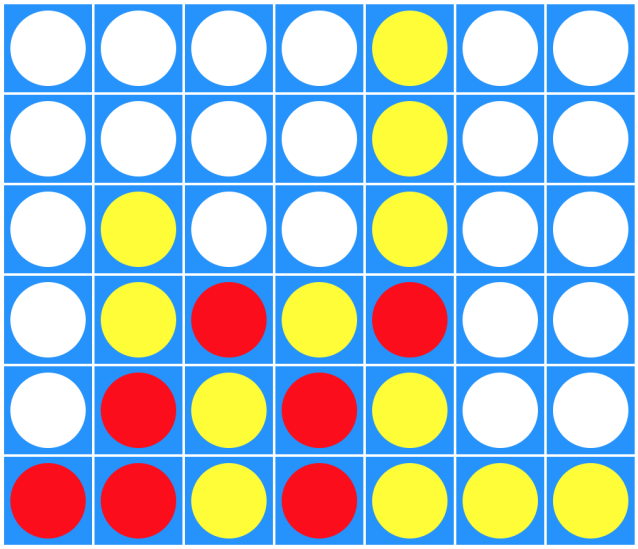


Figure 2.1: Connect Four board without any winners yet

Connect four was mathematically first solved by James Dow Allen (Allen, 1990), and concurrently, independently by Victor Allis (Allis, 1988). Allis (1988) describes a knowledge-based solution with nine strategies. While Allen (1990) discusses multiple different winning strategies. Due to the game's complexity and the computer technology available at the time, only later, the brute-force method could be implemented by Tromp (1995). The name of the brute force artificial intelligence algorithms is minimax or negamax, with optimizations such as alpha-beta pruning. Minimax is a commonly used algorithm for two-player turn-based games. The minimax algorithm is based on searching the tree of possible alternative game states and assigning values to this based on the current player. (Fuller et al., 1973). The alpha-beta pruning makes sure that no branches are explored that will not affect the backed-up value (Fuller et al., 1973). This means no branches are evaluated when there already is a better move discovered. Alpha-Beta pruning makes minimax faster because it does not explore every branch (Fuller et al., 1973). By lowering the depth of the minimax algorithm, the bot will not foresee all situations anymore and play with more errors.

2.4 Technology acceptance model

The Technology Acceptance Model (TAM), depicted in figure Figure 2.2, was firstly proposed by Davis (1985). Afterwards, TAM has widely been adopted to predict user's acceptance of certain information systems or technology (Rattanasampan and Kim, 2002). Using this model we can enable system designers and implementers to test and afterwards evaluate new systems

prior to implementation (Davis, 1985).

TAM examines the role of Perceived Ease of Use (PEOU) and Perceived Usefulness (PU) in their relation between external variables and the probability of actual use, also known as the success of the system (Legris et al., 2003). Variables such as system design characteristics, user characteristics fall into the category "external variables" (Ajzen and Fishbein, 1975)

The TAM does make a couple of assumptions namely, it assumes only voluntary acts of users. Secondly, it assumes technology use only at individual level (Rattanasampan and Kim, 2002).

Next, I will describe all relevant parts of the TAM model below.

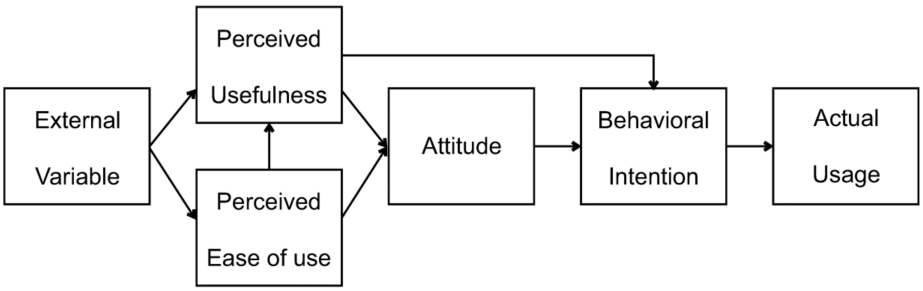


Figure 2.2: Technology Acceptance Model (Davis, 1985)

2.4.1 Perceived Usefulness

The model suggests that Perceived Usefulness (PU), and Perceived Ease of Use (PEOU) are the most important factors to determine system use (Legris et al., 2003). These variables are directly linked and therefore influenced by the external variables. Perceived Usefulness refers to *the degree to which a person believes that using a particular system would enhance his or her job performance*(Davis, 1989). A system that has a high Perceived Usefulness is one where the user be convinced of the existence of a positive use-performance relationship (Davis, 1989).

2.4.2 Perceived Ease of Use

Perceived Ease of Use is described by Davis (1989) as *the degree to which a person believes that using a particular system would be free of effort.* This follows from the idea that "ease" means freedom from difficulty or great effort (Davis, 1989). The TAM claims that an application that is seen to be easier to use than another is more likely to be accepted by users (Davis, 1989).

2.4.3 Attitude Towards Using

Legris et al. (2003) states that external variables intervene indirectly (by first influencing PEOU and PU) influencing attitude towards using. Attitude Towards Using (AT) is defined as *an individual's degree of evaluative affect toward the target behaviour* (Ajzen and Fishbein, 1975). User's overall attitude toward using a given application is hypothesized by the TAM to be a major determinant of the usage (Davis, 1985).

2.4.4 Behavioural Intention to Use

The behavioural Intention to Use (BI) will be influenced by the Attitude Towards Using and Perceived Usefulness indicated by the arrows. Behavioural Intention to Use has been defined by Ajzen and Fishbein (1975) as *an individual's subjective probability that he or she will perform a specified behaviour*. Actual behaviour and actual system use are determined by the intention to perform such a behaviour (Rattanasampan and Kim, 2002).

2.5 Research model and hypotheses

The Technology Acceptance Model was used to base my research model and hypotheses on. The resulting research model is depicted below.

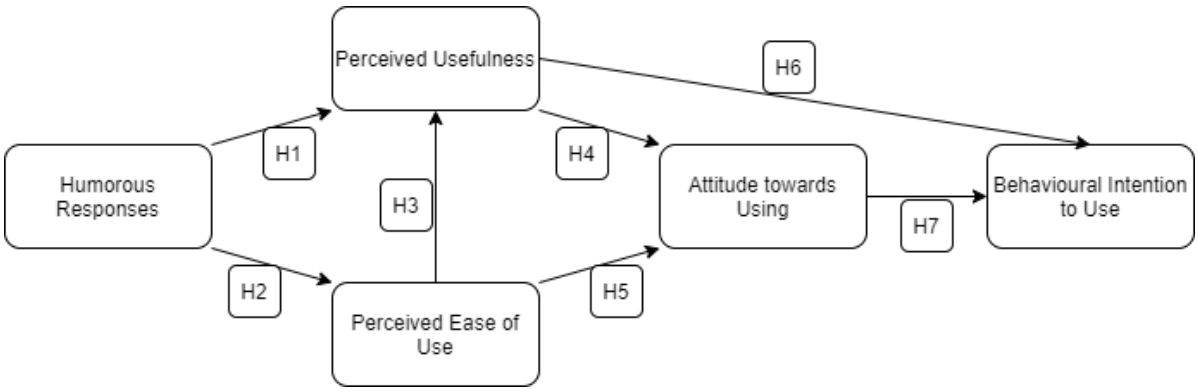


Figure 2.3: Research model with matching hypotheses

The research model results into 7 hypotheses:

H1: A chatbot with humorous responses is perceived as more useful than a chatbot that does not have humorous responses.

H2: A chatbot with humorous responses is perceived as easier to use than a chatbot that does not have humorous responses.

H3: The higher the Perceived Ease of Use is, the higher the perceived Usefulness is.

H4: The higher the Perceived Usefulness is, the higher the Attitude towards Using is.

H5: The higher the Perceived Ease of Use is, the higher the Attitude towards Using is.

H6: The higher the Perceived Usefulness is, the higher the Behavioural Intention to Use is.

H7: The higher the Attitude towards Using is, the higher the Behavioural intention to Use is.

Chapter 3

Method

The Connect Four chatbot and website was programmed together with Britt Deckers and Natalie Hollain. We all had different research questions and experimental conditions within this environment. The control chatbot was used by all of us as baseline condition chatbot.

3.1 Resources

3.1.1 Website Design

For testing the experiment a website was created. To create this website we used HTML, CSS and Javascript files as a basis. These files contained the overall layout and looks as well as most of the implementation of the Connect Four game. Besides that, we used ReactJS, NodeJS and Babel for the front-end and back-end development of our website. To host our website during the testing period we used the pro version of NGROK. This is a service that provides you to host your local host on a public URL domain. Below you will see a visualization of the website when first opened.

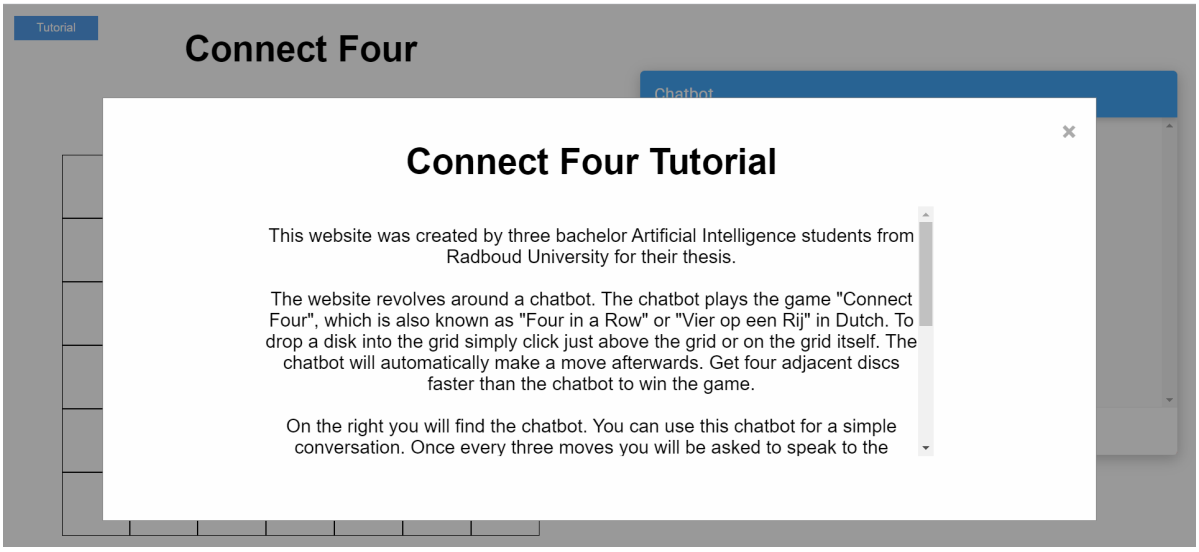


Figure 3.1: Tutorial pop-up

When first opened the website will show a pop-up of the tutorial window. This contains information regarding the use of the website, game and duration. The full text is entailed in Appendix B.1. When closed the website will be visualised as following.

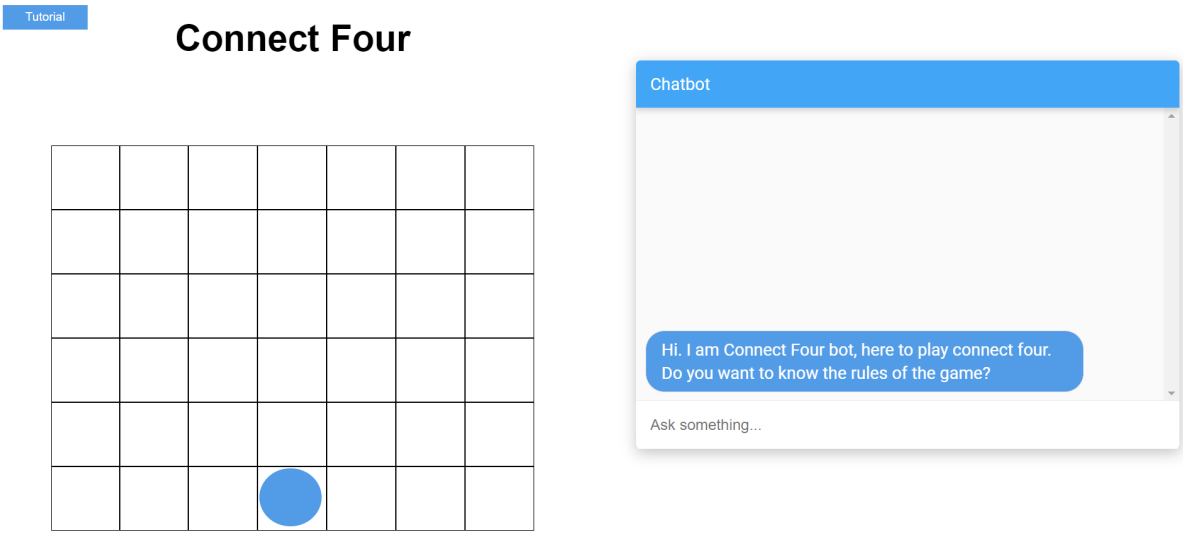


Figure 3.2: Website design

The user is always able to open the tutorial pop-up again by the blue tutorial button in the top left corner.

On the left side of the screen, the simplistic design of the Connect four

game can be seen. When hovering over the grid or just above it, a red (players colour) disc will appear just above that specific column. When clicking the disc will fall as far as possible. When there is a winner the significant winner discs will flash continuously.

On the right, the Dialogflow messenger (a text-based integration of Dialogflow) is visualised. The text bubbles have the appropriate colours based on the colour of that player's discs. The only difference between baseline and humour condition is the text which is sent by the chatbot on the right.

3.1.2 Chatbot

Dialogflow

For most of the small talk dialogue, we used Dialogflow. Dialogflow is a natural language platform used to design and integrate a conversational user interface created by Google.

Dialogflow consists of intents, entities and context. The intents can be created based on certain training phrases which will trigger the machine learning and will randomly choose one of the created responses. Most of the intents are created around frequently asked questions within small talk based on the prebuilt Dialogflow small-talk agent. Examples of this are: asking about the weather, asking about certain hobbies and asking about someone's age or name. There are also intents regarding the game. Such include the standard welcome intent, asking for the rules and asking about board games or Connect four itself. We also created the possibility to ask the chatbot a question or suggestion. The idea is that this made it easier to keep the conversation going when the user didn't know what he/she could talk about.

We chose Dialogflow because of the easy to use interface, machine learning as well as the Webhook possibility to connect with our back-end.

Connect Four integration

There are two types of responses from the chatbot regarding the game of Connect Four. The first one is responding to a specific game question asked by the user. The second one is a response to a certain game state.

Firstly, the user was able to ask the chatbot questions regarding the game. Such questions included *"Who is winning?", "Who's turn is it?"* and *"How much longer do I need to play?"*. These questions (training phrases) were created as intents within Dialogflow but had the webhook connection on for responses. This meant that the responses came from a server.js file via HTTP requests. This was made possible by the use of Node.js. These requests were made possible by ensuring the same port for the server file and the Javascript file. To ensure this sharing of port Babel was used.

The specific responses which the server.js chose were exactly defined in the dialog.json file. Which also ensured the usage of different alike responses.

Secondly, the chatbot made remarks about the game. This was to make the user feel more as if it was playing against a real player. It made remarks such as *"Let me think...", "Looks like I'm winning", "This will be my move", "Looks like you're going to win", "Looks like it ended in a draw"* and *"You made a bad move"*. All these remarks were defined in the dialog.json as well. Alike statements were created for if the player made a good move or there was a specific winner. These statements were triggered by the game Javascript with a certain probability when the game was in a certain state. For example, when the chatbot just made a move it had a 30% chance of saying something similar to *"This will be my move"*. There was also an implementation to make sure the player said something to the chatbot every three moves. When this wasn't done the player wouldn't be able to play disks anymore till it typed something to the chatbot. This was to make sure that the user had some interaction with the chatbot prior to filling in the questionnaire.

Humour condition

The difference between the experimental (humour) condition and the control condition is the added humour. The control condition only contains polite basic emotionless responses. In the humour condition, the basic statements were kept as similar as possible to avoid other causes affecting the results. There were two types of added humour. One relates to joke one-liners. This was previously already successfully done by other researchers as already described in the background. These one-liners were selected from the website by Talmer and Bubble (2013) and each of these was judged as funny by at least 70% of the users. These one-liners were added to all small-talk conversation and some examples can be found in Table 3.1.

Another dialogue was used when responding to the game states. Contradictory it now makes humorous statements which can be found in Table 3.1 as well. All humour used was kept clear of subjects related to ethnicity, sexism, dark subjects, toilet humour and more because on previous research by Niculescu et al. (2013) and Nass and Brave (2005). Below in Figure 3.3 and Figure 3.4, an example is given of the difference between the first couple lines for the control and experiment condition.

Subject	Small talk one-liners	Game state	Responses to game state
Name	Who decided to call it a vet instead of a dogtor?	**Just before the chatbot makes a move**	*pretend to think really hard*
Age	I like having conversations with kids. Grownups never ask me what my third favorite reptile is.	**Just before the chatbot makes a move**	Let me think, Loading almost complete...
Colour	I just found out I'm colorblind. The diagnosis came completely out of the purple	**When the user makes a good move**	Good move, *Crying in the distance*
Weather	Q: What do you call a wet bear? A: A drizzly bear	**When the user wins**	You won RECALIBRATE
Food	Q. What do you call a fake noodle? A. An impasta	**When the chatbot wins**	I won, going to brag to my developers about this
Hobbies	I used to like origami as a hobby. But I gave up as it was a lot of paperwork	**When the user has to say something every 3 moves**	♩ Say something I'm giving up on you ♪ (then you can play again)

Table 3.1: Examples of used humour

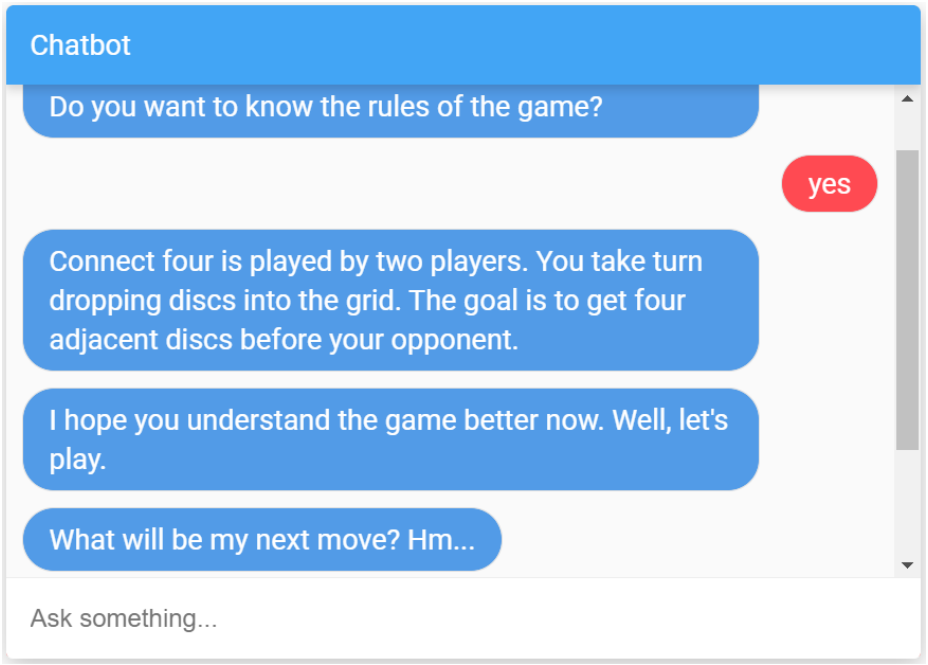


Figure 3.3: Control chat design

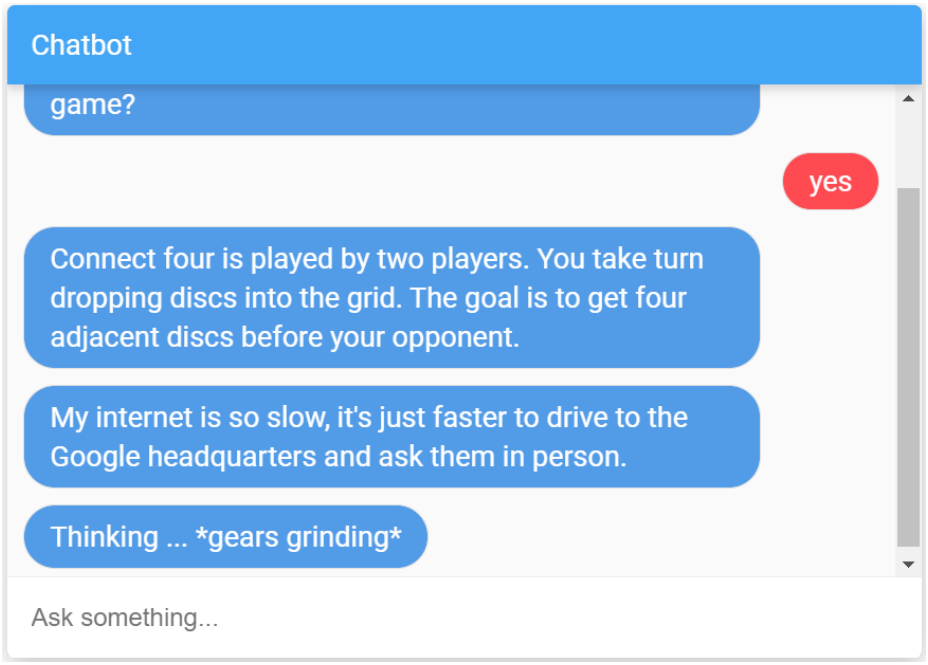


Figure 3.4: Humour chat design

3.1.3 Connect Four

We used Connect four as it is easy to learn and understand. Additionally, the popularity of this game has the advantage that more people already heard of it or know it. At the beginning of the game, the starting player was randomized to account for any disadvantages.

Strategy

For the moves of the chatbot within Connect four, the minimax algorithm was used. On top of this, alpha-beta pruning was used for optimization and computing time. But, we implemented a searching depth of 5 for the tree. This seemed to be challenging enough for a normal player and limited the computing power necessary. To take any frustration into account while testing. There was a part included in the consent form about the chance to either win or lose. As well as a section in the debriefing which apologies for any frustration caused by the difficulty within the game.

Evaluation

To responds accordingly to the moves the player makes, as described in the previous section Connect four integration, an evaluation was implemented. This evaluation compared the moves the player made with the alpha-beta pruning algorithm. When the move was not the best move to possibly be played the chatbot would respond negatively to the move played. The same holds for the expected outcome of the game. When the expected outcome of the game based on the searching tree was towards the player it would tell you it predicts you as a winner. This holds for both players.

3.1.4 Questionnaire

For measuring the user experience we used a questionnaire at the end of the experiment. This questionnaire was created with the software Qualtrics. The questionnaire started with 5 questions related to demographic information such as age, field of work/study and gender. This was used to get an idea of the sample of participants. After this, 4 statements related to previous experience with chatbots, Connect four and technology in total were rated. Hereupon, there were about 5 statements related to each category of the technology acceptance model. Additionally, there were 2 extra statements related to the perceived humour during the conversation. This was added due to the previous observation by Khooshabeh et al. (2011) that when participants perceive the chatbot to be humorous, the chatbot is more effective at socially influencing users. The technology experience statements, TAM related statements as well as perceived humour statements were rated based on the 5-point Likert scale. This method has a scale with the least score (1)

judged as “Strongly disagree” and the highest score (5) judged as “Strongly agree”. Lastly, there were 4 open questions where participants could leave additional information regarding encountered problems, likes and dislikes. Two of those four questions were mandatory to not demand too much from the user. The questionnaire was kept identical in the control and experiment condition to not influence the users’ experience with any other information. Below a summary of the given questions and their related sources. The full questionnaire is included in Appendix C.

Measurement	**Number of Questions**	**Example**	**Source**
Personal Questions	5	”What is your age?”	N\A
Experience Statements	4	”I have experience with chatbots”	N\A
Ease of Use	5	”The chatbot is easy to use”	(Lowry et al., 2012), (Armentano et al., 2015), (Park, 2009), (Van der Heijden, 2004), (Wu and Wang, 2005)
Perceived Usefulness	5	”It is a good idea to use the chatbot to play Connect Four”	(Lowry et al., 2012), (Hassanein and Head, 2007), (Armentano et al., 2015)
Attitude Towards Using	4	”Playing Connect Four with the chatbot is a good idea”	(Armentano et al., 2015), (Park, 2009)
Behavioural Intention	3	”I would use the chatbot frequently if I could”	(Lowry et al., 2012), (Park, 2009), (Wu and Wang, 2005)
Humour	2	”I perceived the chatbot to be humorous”	(Khooshabeh et al., 2011)
Open Questions	4	”What did you like about the chatbot?”	N\A

Figure 3.5: General information about the questionnaire

3.2 Experiment design

This experiment has a between-subject design. This means that each user only gets to test only one of the four experiment conditions (Control, Humour and the two other conditions created by my team members). The goal was to accumulate 20 participants for each condition. This was mostly done via social groups but the experiment was also distributed via Reddit and

SurveyCircle.

The conditions were distributed randomly after accepting the terms disclosed in the consent form and reading the task instructions. This way there was no bias in the distribution of the participants. The full consent form and task instructions can be found in Appendix A and Appendix B respectively. The task instructions indicated overall that you should only have the relating Chrome tab open and not refresh at any moment in the experiment. As well as some instructions to keep any distractions away for full focus. You could contact the researchers if you experienced some difficulties during testing or when the experiment was taking longer than 8 minutes.

After conversing with the chatbot and playing the game Connect four for 5 minutes the link to the questionnaire was provided by the chatbot. At the end of the questionnaire, there was a debriefing text describing the use of the experiment as well as the conditions and where the result may be found afterwards. The full debriefing text can be obtained in Appendix C.4.

In this experiment we only let the participants use the technology once. This makes the experiment subject to the novelty effect. The novelty effect entails that due to the curiosity of the subjects of the new technology they perceived a higher usefulness for the technology (Liu et al., 2009).

The experiment was conducted between May 15th and May 23rd of 2021.

Chapter 4

Results

The Likert scale values which were obtained from the statements in the questionnaire were used to generate these results. Each of the Likert scale scores for the statements was translated into an integer score with 1 being the lowest and 5 the highest. This makes sense because Likert scale data, series of statements that explore different dimensions of a subject, are analyzed at the interval measurement scale (Boone and Boone, 2012). All the scores which belonged together for each category of the TAM or Perceived Humour were combined as a composite score. Each of these resulted in a mean and standard deviation for each aspect. Additional logical data analysis procedures appropriate for interval scale items would include the Pearson's r and ANOVA (Boone and Boone, 2012). For hypotheses 1 and 2 we are going to use the ANOVA one-way test. This is appropriate because we want to analyze variance to test the statistical significance of mean differences between control and experiment (Tabachnick and Fidell, 2007). For the rest of the hypothesis, we will use Pearson's correlation. This test will measure the statistical relationship, or association, between two continuous variables.

4.1 Quantitative results

4.1.1 Outlier removal

As the first step of our statistical process, the outliers were removed. This was done by removing data that was more than 3 standard deviations from the mean. Also known as the z-score method. Because the data set is quite small we only dropped the values of a participant in a certain column instead of the whole participant itself. In the control condition, the Perceived Ease of Use and the Attitude towards Using were replaced with the mean for participant 17. While in the humour condition the Perceived Ease of Use of participant 11 was replaced with the mean.

4.1.2 Participants

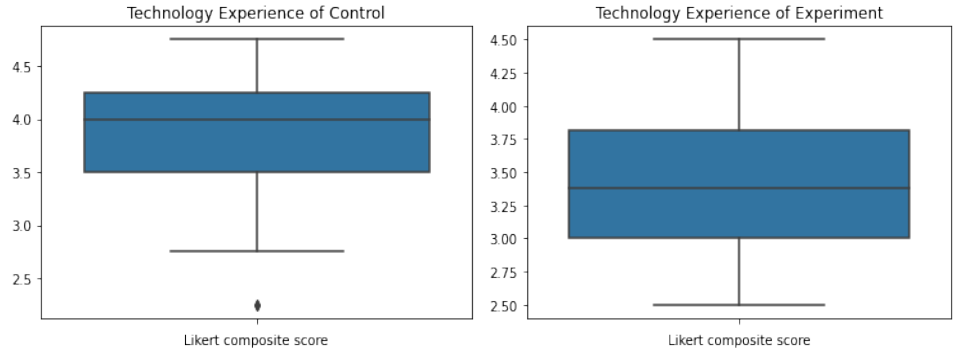
Most of our participants were gathered via social groups as mentioned before. To describe the sample of participants gathered the number of participants, age, gender and nationality were arranged in Table 4.1. A couple of particularities that are noticeable are that a large part of the participants for the humour condition was male. Due to the random assignment of participants, there was a small chance of this happening. Another remark to be made is that most participants for both the control and the humour condition were Dutch. Although there were some other nationalities such as English, German, Belgian, Finnish, Indonesian and Australian.

	Control condition	**Humour condition**
Number of participants	21	20
Average age	23.9	23
Minimum age	19	18
Maximum age	50	51
Gender (Female/Male/Other)	47.6%/52.4%/0%	90%/0.5%/0.5%
Nationality (Dutch/Other)	76.2%/23.8%	75%/25%

Table 4.1: Participant age and gender

In the control condition, 43% said to only have finished secondary school while for the humour condition this was said to be 80%. Most of the other participants either had an HBO (Higher education) diploma or a University bachelors or masters. One participant in the humour condition had a PhD diploma and one mentioned having other. For education, the biggest group of participants were studying Artificial intelligence and Computer science. Making up 43% within the control condition and 45% in the humour condition.

From the 4 questions related to technology experience as described within the questionnaire section the boxplots below in Figure 4.1 could be visualised. The highest-rated statement within this section was the statement *I use technology often*. While the lowest-rated statement overall was *I am an experienced player of Connect Four*. This was the case for the control as well as the humour condition.



(a) Control condition (b) Experiment condition

Figure 4.1: Boxplot of the experience with technology, chatbots and Connect Four

4.1.3 Visualisation

First is an overview given of the measured means and standard deviations for each aspect of TAM and perceived humour in Table 4.2. All means are quite close to each other except for the mean of Perceived Humour which lays higher in the experiment condition.

		Control condition	**Experiment condition**
Perceived Ease of Use	Mean	3.680	3.653
	STD	0.553	0.542
Perceived Usefulness	Mean	3.410	3.40
	STD	0.964	0.795
Attitude towards Using	Mean	3.463	3.588
	STD	0.672	0.708
Behavioural Intention to Use	Mean	2.603	2.733
	STD	0.929	0.928
Perceived Humour	Mean	3.048	3.600
	STD	0.879	0.754

Table 4.2: Means and standard deviations of each aspect of Technology acceptance model including perceived humour

Next up there is a combined boxplot given in Figure 4.2 to show the difference in first quartile, the median and the third quartile between control and experiment condition.

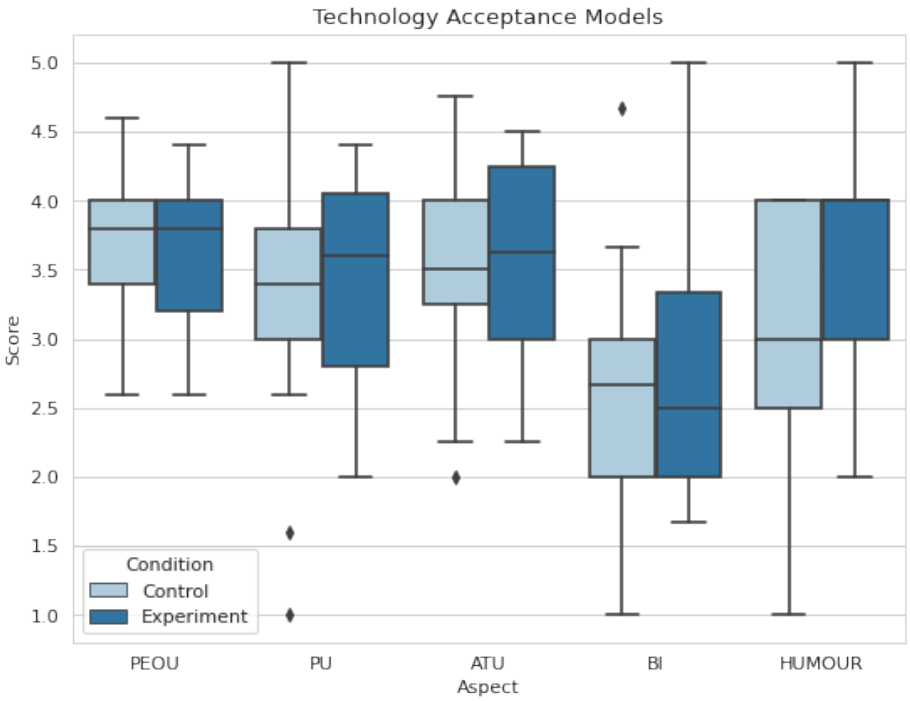


Figure 4.2: The boxplots of the aspects of the Technology acceptance model

Lastly, the Likert histograms of the control and experiment condition are given in Figure 4.3 and 4.4. This is the actual data on which all means and standard deviations are based.

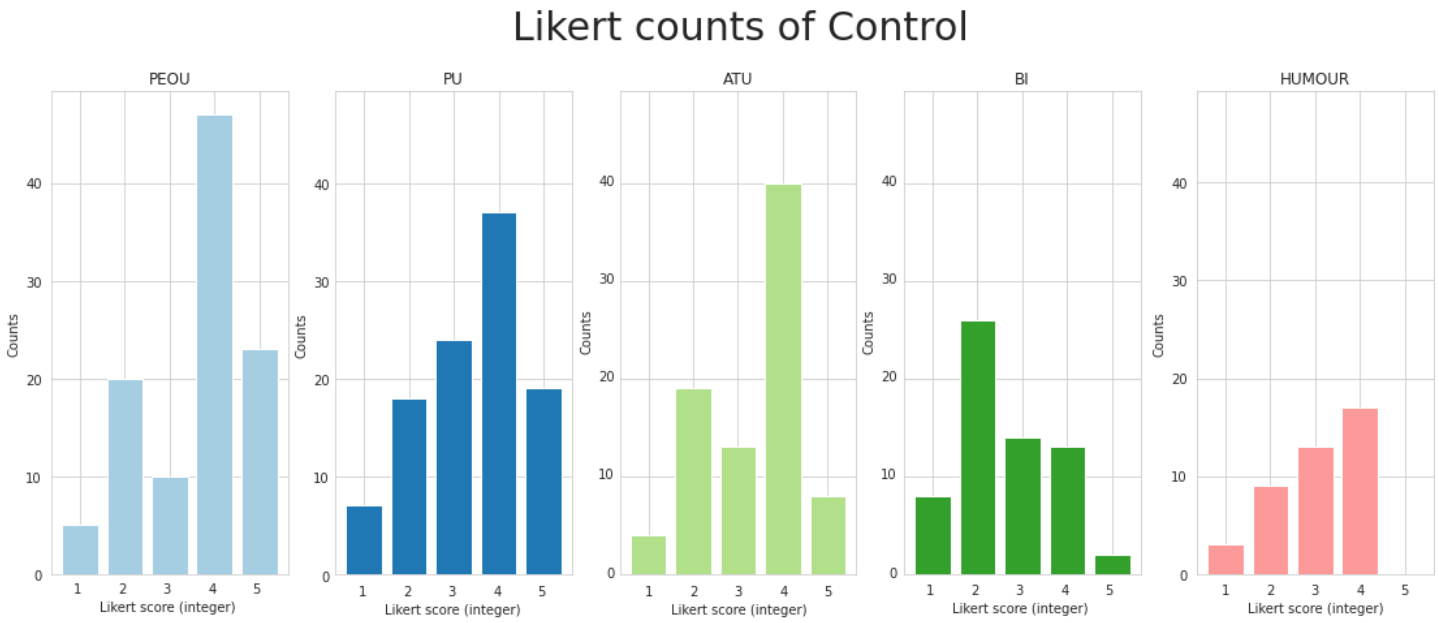


Figure 4.3: Likert count of the aspects within the control condition

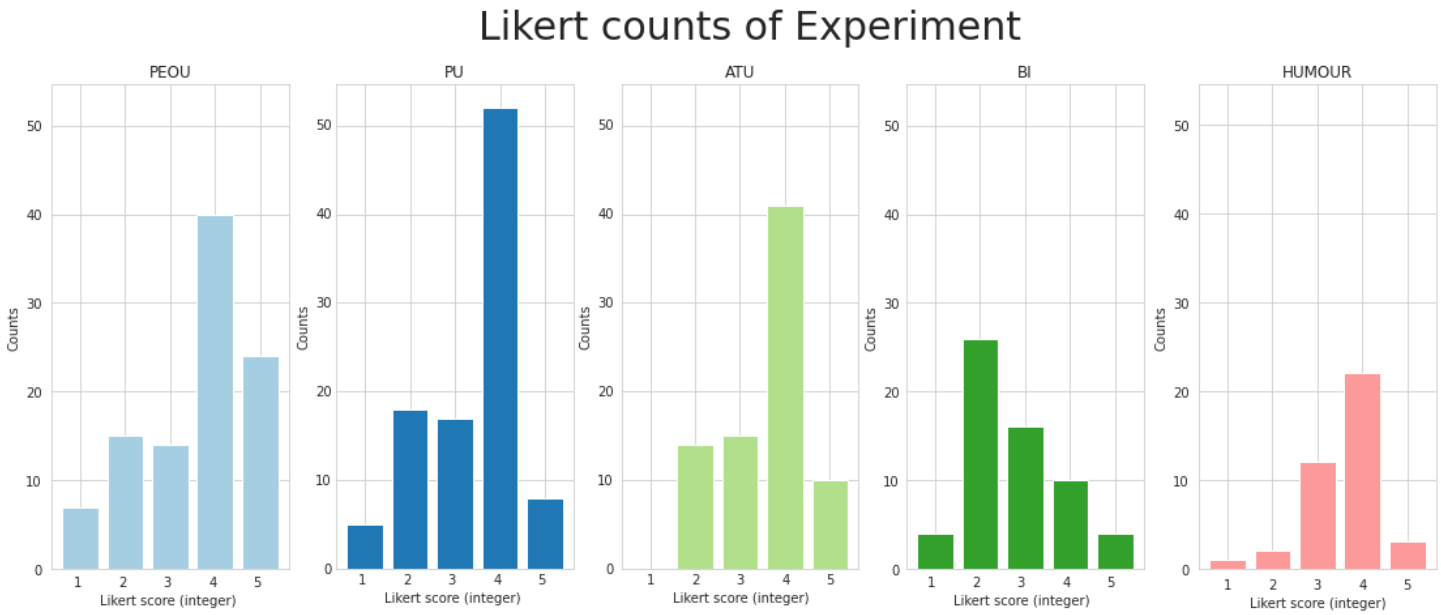


Figure 4.4: Likert count of the aspects within the humour condition

4.1.4 ANOVA

For the following two hypotheses we will be using ANOVA as the statistical method.

H1: A chatbot with humorous responses is perceived as more useful than a chatbot that does not have humorous responses.

H2: A chatbot with humorous responses is perceived as easier to use than a chatbot that does not have humorous responses.

ANOVA makes a couple of assumptions. These include normality, equal variance and independence of the data. To measure normality we use the Shapiro-Wilk test. This resulted in the Perceived Usefulness within the experiment condition and both Perceived Humour data not being normally distributed (had a p-value smaller than 0.05). Next up to test for equal variance we use Levene's test. This one suggests that all data has similar variance (had p-values bigger than 0.05). Because of the use of a between-subject design of the experiment. All participants only encountered one condition. Therefore independence was assumed.

Although a couple of normality assumptions are violated, Norman (2010) suggests this may form no problem since such a statistical test is robust.

	F-value	**p-value**
Perceived Ease of Use	0.026	0.874
Perceived Usefulness	0.001	0.973
Perceived Humour	4.643	0.037

Table 4.3: ANOVA results

The p-values for the relevant ANOVA tests are portrayed in Table 4.3. The p-values of Perceived Ease of Use and Perceived Usefulness are both higher than 0.05. Therefore, both hypotheses 1 and 2 can be rejected. The p-value of perceived humour is lower than 0.05. This means that there is a significant difference between the perceived humour for control and experiment condition.

4.1.5 Pearson's correlation coefficient

To determine the correlation between different aspects of the TAM model Pearson's correlation coefficient was used. These resulting p-values were used to determine the significance for the rest of the hypotheses given below.

H3: The higher the Perceived Ease of Use is, the higher the Perceived Usefulness is.

H4: The higher the Perceived Usefulness, the higher the Perceived Usefulness is.

H5: The higher the Perceived Ease of Use is, the higher the Attitude Towards Using is.

H6: The higher the Perceived Usefulness, the higher the Behavioural Intention to Use is.

H7: The higher the Attitude Towards Using is, the higher the Behavioural Intention to Use is.

Following Figure 4.5 and 4.6 is the visualisation of the given Pearson's correlation coefficient in Table 4.4 for the control and experiment condition. Each of these plots corresponds to one of the hypothesis.

	Perceived Ease of Use	**Perceived Usefulness**	**Attitude towards Using**	**Behavioral Intention to Use**
Perceived Ease of Use	1			
Perceived Usefulness	0.457	1		
Attitude towards Using	0.508	0.676	1	
Behavioral Intention to Use	0.392	0.657	0.712	1

Table 4.4: Pearson's correlation coefficient results for the combined data

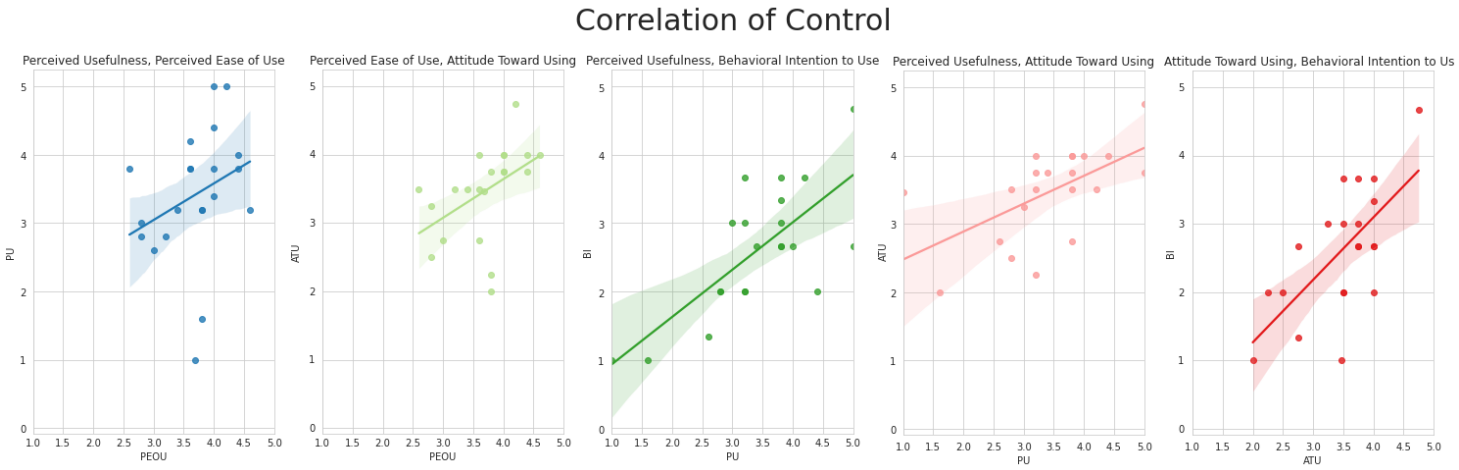


Figure 4.5: The boxplots of the aspects of the Technology acceptance model

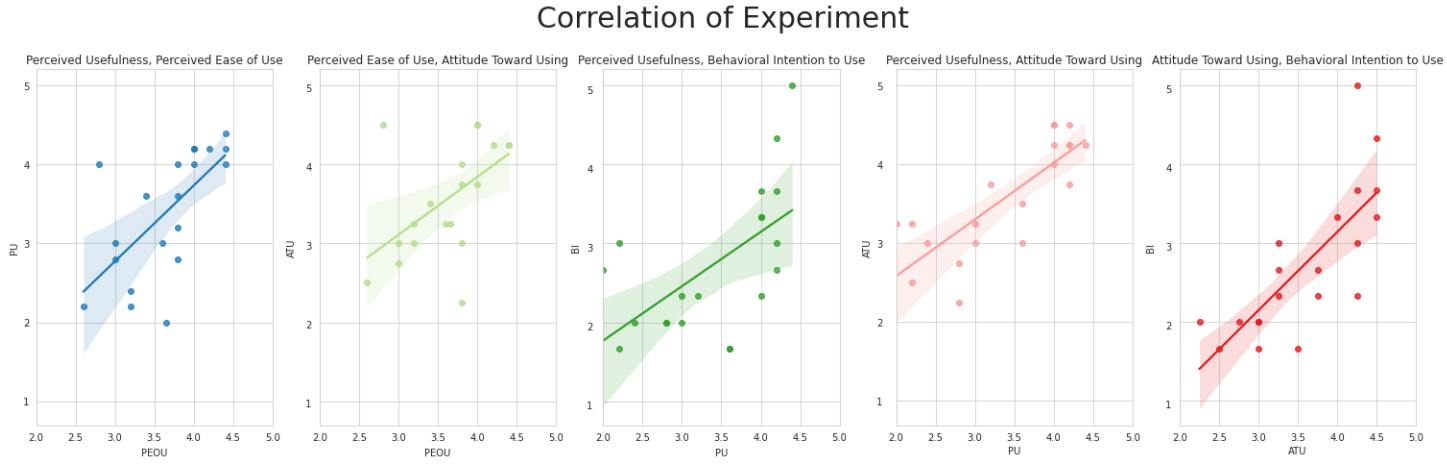


Figure 4.6: The boxplots of the aspects of the Technology acceptance model

Given that all p-values were below 0.05 for the hypotheses, all of the hypotheses can be accepted.

Another interesting factor to notice is that the Pearson's correlation p-value for Perceived Humour and Perceived Ease of Use was also below 0.05. Based on Figure 4.7 there is a significant positive correlation between Perceived humour and Perceived Ease of Use. The p-value between Perceived Humour and Perceived Usefulness was just above the significance level namely 0.082.

	Perceived Ease of Use	Perceived Usefulness	Attitude towards Using	Behavioral Intention to Use	Perceived Humour
Perceived Humour	0.392	0.275	0.123	0.238	1

Table 4.5: Pearson's correlation coefficient results for the Perceived Humour

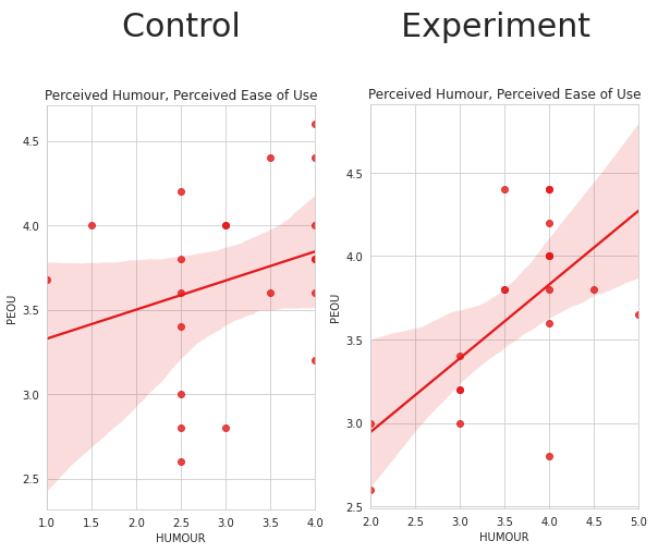


Figure 4.7: The correlation between Perceived Ease of Use and Perceived Humour for control and experiment condition

4.2 Qualitative results

Below some frequently mentioned remarks are noted. This data comes from the open questions at the end of the questionnaire.

- Quite a lot of users mentioned the limited dialogue of both chatbots. Either the chatbot was too repetitive or they found it to not have a large enough vocabulary.
- In both conditions a few participants noted the chatbot as human-like and the conversation being enjoyable. They mentioned the feeling as if they had a real opponent and feel challenged to do more. One control participant even mentioned that it helped with their social anxiety to talk with a chatbot.
- The rule that you have to say something every 3 turns was mostly

perceived as annoying but some said it was a great way of keeping the conversation going.

- Within the control condition some participants mentioned it made jokes or witty comments and was found to be funny.
- Some participants of the humour condition noted that the jokes were not funny or interpreted them as rude. Although most mentioned it to be funny.
- In the humour condition someone noted that it doesn't respond well to humour made by the user.

Chapter 5

Discussion

The issues that chatbots confront are for example being irritating, repetitive, or intimidating. These difficulties could be overcome by humour because humans use humour in a similar manner (Binsted, 1995). Previous research in the field of chatbots already suggests that adopting humour in the conversation improves likeability and may contribute to a better user experience (Niculescu and Banchs, 2019). In this thesis, we gain insight into the difference in user experience between the humorous condition and control condition of a chatbot within a gaming environment. To test the effect, a humorous chatbot and control chatbot were created. The humorous chatbot had additional one-liner jokes for all the small talk and humorous remarks regarding Connect Four. The experiment was designed as a between-subjects study, with the user experience being analyzed using a questionnaire based on the Technology Acceptance Model.

Surprisingly, hypotheses 1 and 2 have to be rejected. This was the case as the means, of the control and humour condition, for the measured concepts Perceived Usefulness and Perceived Ease of Use didn't differ significantly. This means there is not a perceived difference for the control and humour condition which is not in line with previous work. This observation could be for the reason that humour is highly shaped by culture (Ruch, 2007) and there were in fact multiple different nationalities participating in the study. This does not explain the whole insignificance because the biggest part of the participants was Dutch. As well as the personal nature of humour why it can often go unremarked, misunderstood or perceived as offensive (Reimann et al., 2010). Another remark is that quite a lot of participants in the control indicated the chatbot to be funny, which could be relating to participant bias (Greenberg et al., 1969), and some indicated the humorous condition to not be funny. This can be seen in Figure 4.3 and Figure 4.4. Therefore it could have influenced the Perceived Usefulness and Perceived Ease of Use of the control positively and of the experiment negatively. And in fact, there was a significant correlation (p-value of 0.011) found between Perceived Humour

and Perceived Ease of Use and a non-significant p-value of 0.082 between Perceived Humour and Perceived Usefulness. This not completely a new finding. Khooshabeh et al. (2011) discovered that a humorous virtual agent is more effective at socially influencing people when it is viewed as funny. The perceived humour could have a bigger influence on the user experience than originally thought. Which in a way makes sense, if you don't experience or recognise the humour used in the conversation how would it be able to affect your user experience overall. This might be an interesting model to test for future research in this field.

There was a significant difference between perceived humour for the control and experiment condition. This means that the participants noticed and indicated clearly the difference between the humorous and the non-humorous conversation. This makes the forms of humour that were used useful and this is in line with the forms of humour used in previous research. There could also be different types of humour implemented which will create a similar effect. For example the approach of Blinov et al. (2017), they use funny Tweet generation instead of jokes. Another approach could be the generation of internet memes (*a piece of content spreading online from user to user and changing along the way*) which has humour as its key component (Börzsei, 2013). A whole different approach is making a chatbot able to recognize humour expressed by the user (Augello et al., 2008) (Pilato et al., 2008).

All hypotheses from H4 onward relating to the Technology Acceptance model are all to be accepted. This was tested by Pearson's correlation coefficient which gave a p-value below 0.05 for all hypotheses. This makes sense because the model has been validated in several empirical studies, and the tools used to implement the model have proven to be of high quality and give statistically accurate findings (Olushola and Abiola, 2017).

Due to COVID-19, the whole designing and testing process has been conducted online. This also means that during the testing phase we can not assure that there was the full focus from the participants on the task at hand. This means that it could be that the experimental conditions were not equal for all participants. The experiment could be conducted again in a sterile environment for focus and for visual indicators such as smiles and laughs which were seen by Morkes et al. (1998). Also because of online communication between team members, there may be more limitations than expected in the end-product.

Most of the users were gathered through social circles. This made the sample average age for both experiment and control quite young, namely 23.9 and 23. Also because of random assignment of the participants over the multiple conditions, the humour condition consisted of 90% male participants. These factors can both have an influence on the rated conversational quality. Namely, Shah et al. (2016) mentions that different demographic

groups tend to rate chatbots differently. Specifically, conversations were rated higher by younger and female users, respectively. The nationality of both the control condition and the humour condition were mostly Dutch. This probably is also due to the gathering of participants through social circles. This could have an effect on the understood humour and the way the concept perceived humour was distributed. Niculescu and Banchs (2019) describes how understanding the humorous intention is enhanced by having a good command of the language, namely English, used in interaction with the agent.

All the testing was only done once by each participant. This made the experiment subject to the novelty effect. A set-up where the chatbot would be tested twice could potentially decrease the novelty effect. Still, overcoming the novelty effect remains a large difficulty. According to Prochaska and DiClemente (1982), it might take up to six months for a new behaviour to become a habit, such as full adoption of new technology.

In this thesis, we worked with a chatbot that played Connect Four while having a conversation with you. Due to the fact that some subjects indicated that they were not experienced in Connect Four, the quite high difficulty of the game, as well as that it was a gaming environment overall could have affected the user experience. This makes that the results can not be generally accepted for all different kind of chatbots. It might be interesting to see if the results generalize for different gaming environments such as chess. Also having varying levels of difficulty for the game of Connect Four and seeing the effects of this on the user experience.

Chapter 6

Conclusions

In this thesis, insight was gained into the role humour plays in the user experience within a gaming environment. The results suggest that the role of humour, in form of one-liner jokes and remarks about the game state, does not have a significant effect on the Perceived Ease of Use and Perceived Usefulness. This is not corresponding to previous research within the field of humour and chatbots. The cause for this could be related to the gaming environment, the limited vocabulary, the small sample size or the representation of participants. The findings that all relations within the Technology Acceptance Model are significant is in line with previous work on this model. Future work within this area could focus on a more specific environment, a bigger, more representative sample size, more foul proof dialogue and other implementations of humour.

Chapter 7

Acknowledgements

First and foremost, I want to express my gratitude to Marianna De Sá Siqueira for serving as my thesis supervisor. Second, I would like to thank Britt Deckers and Natalie Hollain for their assistance with the chatbot's development, the questionnaire, and the analysis of the data. Finally, I want to express my appreciation to Tibor Bosse for serving as my thesis's second reader. Lastly, I would like to thank all the participants in the experiment.

Bibliography

Adams, D. (1995). *The hitch hiker's guide to the galaxy: a trilogy in five parts*. Random House.

Ajzen, I. and Fishbein, M. (1975). A bayesian analysis of attribution processes. *Psychological bulletin*, 82(2):261.

Allen, J. D. (1990). Expert play in connect-four.

Allis, V. (1988). A knowledge-based approach of connect-four. the game is solved: White wins.

Armentano, M. G., Christensen, I., and Schiaffino, S. (2015). Applying the technology acceptance model to evaluation of recommender systems. *Polibits*, (51):73–79.

Augello, A., Gambino, O., Cannella, V., Pirrone, R., Gaglio, S., and Pilato, G. (2011). An emotional talking head for a humoristic chatbot. *Applications of Digital Signal Processing*, 319.

Augello, A., Saccone, G., Gaglio, S., and Pilato, G. (2008). Humorist bot: Bringing computational humour in a chat-bot system. In *2008 International Conference on Complex, Intelligent and Software Intensive Systems*, pages 703–708.

Babu, S., Schmugge, S., Barnes, T., and Hodges, L. F. (2006). "what would you like to talk about?" an evaluation of social conversations with a virtual receptionist. In *International Workshop on Intelligent Virtual Agents*, pages 169–180. Springer.

Baron, R. A. and Ball, R. L. (1974). The aggression-inhibiting influence of nonhostile humor. *Journal of Experimental Social Psychology*, 10(1):23–33.

Binsted, K. (1995). Using humour to make natural language interfaces more friendly. In *Proceedings of the AI, ALife and Entertainment Workshop, Intern. Joint Conf. on Artificial Intelligence*.

Blinov, V., Mishchenko, K., Bolotova, V., and Braslavski, P. (2017). A pinch of humor for short-text conversation: An information retrieval approach. In Jones, G. J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 3–15, Cham. Springer International Publishing.

Boone, H. N. and Boone, D. A. (2012). Analyzing likert data. *Journal of extension*, 50(2):1–5.

Börzsei, L. K. (2013). Makes a meme instead. *The Selected Works of Linda Börzsei*, pages 1–28.

Bradeško, L. and Mladenić, D. (2012). A survey of chatbot systems through a loebner prize competition. In *Proceedings of Slovenian language technologies society eighth conference of language technologies*, pages 34–37. Institut Jožef Stefan Ljubljana, Slovenia.

Brandtzaeg, P. B. and Følstad, A. (2017). Why people use chatbots. In Kompatsiaris, I., Cave, J., Satsiou, A., Carle, G., Passani, A., Kontopoulos, E., Diplaris, S., and McMillan, D., editors, *Internet Science*, pages 377–392, Cham. Springer International Publishing.

Bricker, V. R. (1980). The function of humor in zinacantan. *Journal of Anthropological Research*, 36(4):411–418.

Castell, P. J. and Goldstein, J. H. (1977). Social occasions for joking: a cross-cultural study. In *It's a funny thing, humour: International conference on humour and laughter*, pages 193–197.

Chaves, A. P. and Gerosa, M. A. (2019). How should my chatbot interact? A survey on human-chatbot interaction design. *CoRR*, abs/1904.02743.

Civikly, J. (1983). A comparison of male and female uses of humor types and humor functions. In *annual meeting of the International Communication Association, Dallas, TX*.

Civikly, J. (1989). Humor and teaching: The first years of teaching. In *annual meeting of the Speech Communication Association, San Francisco, CA*.

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. PhD thesis, Massachusetts Institute of Technology.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340.

D'onfro, J. (2015). Microsoft created a chatbot in china that has millions of loyal followers who talk to it like in the movie 'her'.

Driessen, H. (2015). Humor, anthropology of. *Wright, JD (ed.), International encyclopedia of the social & behavioral sciences (2nd ed.)*, pages 416–419.

Dybala, P., Ptaszynski, M., Higuchi, S., Rzepka, R., and Araki, K. (2008). Humor prevails!-implementing a joke generator into a conversational system. In *Australasian Joint Conference on Artificial Intelligence*, pages 214–225. Springer.

Dybala, P., Ptaszynski, M., Rzepka, R., and Araki, K. (2009). Humoroids: conversational agents that induce positive emotions with humor. In *AAMAS'09 Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, volume 2, pages 1171–1172. ACM.

Erdodi, L. and Lajiness-O'Neill, R. (2012). Humor perception in bilinguals: Is language more than a code? *Humor*, 25(4):459–468.

Følstad, A. and Brandtzæg, P. B. (2017). Chatbots and the new world of hci. *interactions*, 24(4):38–42.

Fuller, S. H., Gaschnig, J. G., Gillogly, J., et al. (1973). *Analysis of the alpha-beta pruning algorithm*. Department of Computer Science, Carnegie-Mellon University.

Gibson, D. E. (1994). Humor consulting: Laughs for power and profit in organizations.

Goldstein, J. H., Silverman, A., and Anderson, P. (1976). A cross-cultural investigation of humour: Appreciation of jokes varying in familiarity and hostility. In *XXIst International Congress of Psychology, Paris*.

Graham, E. E., Papa, M. J., and Brooks, G. P. (1992). Functions of humor in conversation: Conceptualization and measurement. *Western Journal of Communication (Includes Communication Reports)*, 56(2):161–183.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., and Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64(326):520–539.

Hassanein, K. and Head, M. (2007). Manipulating perceived social presence through the web interface and its impact on attitude towards online shopping. *International Journal of Human-Computer Studies*, 65(8):689–708.

Huan, C. and Szafir, D. (2001). No joke: examining the use of humor in computer-mediated learning. *Unpublished material*.

Khooshabeh, P., McCall, C., Gandhe, S., Gratch, J., and Blascovich, J. (2011). Does it matter if a computer jokes. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 77–86.

Landy, D. and Mettee, D. (1969). Evaluation of an aggressor as a function of exposure to cartoon humor. *Journal of Personality and Social Psychology*, 12(1):66.

Legris, P., Ingham, J., and Collerette, P. (2003). Why do people use information technology? a critical review of the technology acceptance model. *Information Management*, 40(3):191–204.

Linstead, S. (1985). Jokers wild: The importance of humour in the maintenance of organizational culture. *The Sociological Review*, 33(4):741–767.

Liu, S.-H., Liao, H.-L., and Pratt, J. A. (2009). Impact of media richness and flow on e-learning technology acceptance. *Computers & Education*, 52(3):599–607.

Lowry, P. B., Gaskin, J., Twyman, N., Hammer, B., and Roberts, T. (2012). Taking 'fun and games' seriously: Proposing the hedonic-motivation system adoption model (hmsam). *Journal of the association for information systems*, 14(11):617–671.

Martin, G. N. and Sullivan, E. (2013). Sense of humor across cultures: A comparison of british, australian and american respondents. *North American Journal of Psychology*, 15(2).

Martin, R. A. and Ford, T. (2018). *The psychology of humor: An integrative approach*. Academic press.

Medhi Thies, I., Menon, N., Magapu, S., Subramony, M., and O'Neill, J. (2017). How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians. In *Human-Computer Interaction - INTERACT 2017*, pages 441–459, Cham. Springer International Publishing.

Morkes, J., Kernal, H. K., and Nass, C. (1998). Humor in task-oriented computer-mediated communication and human-computer interaction. In *CHI 98 Conference Summary on Human Factors in Computing Systems*, pages 215–216.

Nass, C., Fogg, B., and Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, 45(6):669–678.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103.

Nass, C., Moon, Y., and Carney, P. (1999). Are people polite to computers? responses to computer-based interviewing systems 1. *Journal of applied social psychology*, 29(5):1093–1109.

Nass, C. I. and Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, MA.

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., and See, S. L. (2013). Making social robots more attractive: the effects of voice pitch, humor and empathy. *International journal of social robotics*, 5(2):171–191.

Niculescu, A. I. and Banchs, R. E. (2015). Strategies to cope with errors in human-machine spoken interactions: using chatbots as back-off mechanism for task-oriented dialogues. *Proceedings of ERRARE, Sinaia, Romania*.

Niculescu, A. I. and Banchs, R. E. (2019). Humor intelligence for virtual agents. In *9th International Workshop on Spoken Dialogue System Technology*, pages 285–297. Springer.

Nijholt, A. (2007). Conversational agents and the construction of humorous acts. In *Conversational Informatics: An Engineering Approach*, pages 21–47. Wiley.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632.

Olushola, T. and Abiola, J. (2017). The efficacy of technology acceptance model: A review of applicable theoretical models in information technology researches. *Journal of Research in Business and Management*, 4(11):70–83.

Park, S. Y. (2009). An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. *Journal of Educational Technology & Society*, 12(3):150–162.

Pilato, G., Augello, A., Vassallo, G., and Gaglio, S. (2008). Ehebby: An evocative humorist chat-bot. *Mobile Information Systems*, 4(3):165–181.

Prochaska, J. O. and DiClemente, C. C. (1982). Transtheoretical therapy: toward a more integrative model of change. *Psychotherapy: theory, research & practice*, 19(3):276.

Ptaszynski, M., Dybala, P., Higuhi, S., Shi, W., Rzepka, R., and Araki, K. (2010). Towards socialized machines: Emotions and sense of humour in conversational agents. *Web intelligence and intelligent agents*, 173.

Rattanasampan, W. and Kim, S. (2002). A framework to study technology use: Alternatives to technology acceptance model. *AMCIS 2002 Proceedings*, page 127.

Reeves, B. and Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press Cambridge, UK.

Reimann, A. et al. (2010). Intercultural communication and the essence of humour. *Journal of the Faculty of International Studies, Utsunomiya University*, (29):23–34.

Roy, D. F. (1959). " banana time" job satisfaction and informal interaction. *Human organization*, 18(4):158–168.

Ruch, W. (2007). *The sense of humor: Explorations of a personality characteristic*. de Gruyter.

Shah, H., Warwick, K., Vallverdú, J., and Wu, D. (2016). Can machines talk? comparison of eliza with modern dialogue systems. *Computers in Human Behavior*, 58:278–295.

Smith, C. M. and Powell, L. (1988). The use of disparaging humor by group leaders. *Southern Speech Communication Journal*, 53(3):279–292.

Smith, R. E., Ascough, J. C., Ettinger, R. F., and Nelson, D. A. (1971). Humor, anxiety, and task performance. *Journal of Personality and Social Psychology*, 19(2):243.

Stocking, S. H. and Zillmann, D. (1976). Effects of humorous disparagement of self, friend, and enemy. *Psychological Reports*, 39(2):455–461.

Sykes, A. J. (1966). Joking relationships in an industrial setting. *American anthropologist*, 68(1):188–193.

Tabachnick, B. G. and Fidell, L. S. (2007). *Experimental designs using ANOVA*. Thomson/Brooks/Cole Belmont, CA.

Talmer and Bubble (2013). 4603 funny one liners - funniest short jokes.

Tromp, J. (1995). John's connect four playground.

Tseng, Y.-H., Hsu, W.-L., Wu, W.-S., Gu, Y.-C., and Chen, H.-C. (2020). Implementation and evaluation of a retrieval-based chinese humor chatbot. *Journal of Library & Information Studies*, 18(2).

Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, LIX(236):433–460.

Van der Heijden, H. (2004). User acceptance of hedonic information systems. *MIS quarterly*, pages 695–704.

Weaver, J., Zillmann, D., and Bryant, J. (1988). Effects of humorous distortions on children's learning from educational television: Further evidence. *Communication Education*, 37(3):181–187.

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

Whitaker, C. A. (1975). Psychotherapy of the absurd: With a special emphasis on the psychotherapy of aggression. *Family Process*, 14(1):1–16.

Wu, J.-H. and Wang, S.-C. (2005). What drives mobile commerce?: An empirical evaluation of the revised technology acceptance model. *Information & management*, 42(5):719–729.

Xu, A., Liu, Z., Guo, Y., Sinha, V., and Akkiraju, R. (2017). A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510.

Zillmann, D. and Bryant, J. (1983). Uses and effects of humor in educational ventures.

Appendix A

Informed consent form

Consent Form

Information

Purpose of the Study
This study is conducted within the Bachelor Thesis of the Artificial Intelligence Programme of the Radboud University.
The study involves a chatbot that plays a simple board game. You will converse and play with the chatbot. You may win or lose the game. Afterwards, you will fill out a questionnaire.

Voluntary Participation
Your participation is voluntary. During the study you can indicate at any moment that you want to quit participating, without having to explain why you want to quit. You can leave the study at any moment if you want.

Anonymity
The collected information shall be processed anonymously. This means that the results cannot be traced back to you later on. We therefore cannot inform you about your personal results after the study has been completed. However, we could inform you about the results of the study as a whole. If you wish to be informed about the results of this study, please let us know.

Data Usage
Your data can be used as a part of data sets, articles and presentations. The anonymized research data is accessible to other researchers for a period of at least 10 years. When we share data with other researchers, these data cannot be traced back to you. The data we collect entails the chat history with the chatbot and questionnaire data.

Contact
You can contact the researchers about further questions through the following email address: b.deckers@student.ru.nl

Consent
By consenting, you agree with the following statements:

- The aim of the research study has been outlined to me.
- I was given the opportunity to ask questions regarding the research study.
- I participate voluntarily in the research study.
- I understand that I can stop at any point during the research study, should I wish to do so.
- I understand how the data of the research study will be stored and how they will be used.
- I consent to participating in the research study.

By clicking on 'I consent' below, you consent to participating in this study.

◯ I consent
◯ I do NOT consent

Appendix B

Task Instructions

You are going to play Connect Four (also known as 'Four in a Row', in Dutch: 'Vier op een Rij') with the chatbot for about 5 minutes. The chatbot will have a conversation with you while you play. The link to the chatbot is provided after the task instructions. Please read the following points **carefully**.

1. Please make sure you are running the chatbot in Google Chrome. Other browsers are not compatible. If you are currently using another browser to read the instructions, open the chatbot link provided to you in Google Chrome instead.

2. Use your computer or laptop for this session. Mobile phones and other devices are not supported.

3. Keep this tab open at all times. Do **NOT** continue to the next page of this tab until you are done with the chatbot.

4. Please make sure your environment is distraction-free. Close any other programs on your computer, interact with no other persons and put your phone on silent mode.

5. Interact with the chatbot until you receive a code for the questionnaire. Do **NOT** reload the page unless you stop receiving messages from the chatbot for several minutes, before you have received the link to the questionnaire. If you stop receiving messages before the end of the session, a server error has occurred. Please only reload the chatbot page in this case.

6. Do **NOT** close the chatbot window during the session.

7. Open any link provided on this page in a separate tab.

8. If it takes more than 8 minutes from the start of the session for you to receive the link for the questionnaire, please contact the researchers.

B.1 Website Tutorial Text

This website was created by three bachelor Artificial Intelligence students from Radboud University for their thesis.

The website revolves around a chatbot. The chatbot plays the game "Connect Four", which is also known as "Four in a Row" or "Vier op een Rij" in Dutch. To drop a disk into the grid simply click just above the grid or on the grid itself. The chatbot will automatically make a move afterwards. Get four adjacent discs faster than the chatbot to win the game.

On the right you will find the chatbot. You can use this chatbot for a simple conversation. Once every three moves you will be asked to speak to the chatbot before continuing the game. If you cannot make a move, try to talk to the chatbot first. The chatbot can always tell you the rules again simply by asking.

You can also ask for suggestions for topics. Please try to speak to the chatbot often.

This tutorial can be reopened with the button in the top-left corner of the screen.

The play session will take about 5 minutes. At the end you will be asked to fill out a questionnaire. This would greatly help out with our thesis.

Appendix C

Questionnaire

C.1 General questions

What is your age? _____

What is your gender?

◯ Female
◯ Male
◯ Other

What is your nationality?

◯ Dutch
◯ German
◯ English
◯ French
◯ Other: _____

What is the highest level of education that you have finished?

◯ None
◯ Primary school
◯ Secondary school
◯ Higher Education (HBO) Bachelor
◯ Higher Education (HBO) Master
◯ University (WO) Bachelor
◯ University (WO) Master
◯ PhD
◯ Other

In which field do you work or study?

◯ Artificial Intelligence
◯ Computer Science
◯ Psychology
◯ Other: _____

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I am knowledgeable about technology	◯	◯	◯	◯	◯
I use technology often	◯	◯	◯	◯	◯
Playing against the chatbot was enjoyable	◯	◯	◯	◯	◯
I have experience with chatbots	◯	◯	◯	◯	◯
I am an experienced player of Connect Four	◯	◯	◯	◯	◯

Table C.1: Technology Experience

C.2 TAM-related questions

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
The chatbot is easy to use	○	○	○	○	○
Learning how to use the chatbot was easy	○	○	○	○	○
Interacting with the chatbot is clear and understandable	○	○	○	○	○
By using the chatbot, it is easier to play Connect Four	○	○	○	○	○
Using the chatbot does not require a lot of mental effort	○	○	○	○	○

Table C.2: Perceived Ease of Use

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Using the chatbot helps me feel like I am playing against an opponent	○	○	○	○	○
It is a good idea to use the chatbot to play Connect Four	○	○	○	○	○
The chatbot helped me better pass time	○	○	○	○	○
The chatbot provides a quality gaming experience	○	○	○	○	○
The chatbot is useful for playing a game and chatting	○	○	○	○	○

Table C.3: Perceived Usefulness

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Using the chatbot makes it more interesting to play Connect Four	◯	◯	◯	◯	◯
I would like to play Connect Four with the chatbot again	◯	◯	◯	◯	◯
Using the chatbot to play a game and have a chat seems fun	◯	◯	◯	◯	◯
Playing Connect Four with the chatbot is a good idea	◯	◯	◯	◯	◯

Table C.4: Attitude Toward Using

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Assuming I had access to the chatbot in the future, I intend to use it	◯	◯	◯	◯	◯
Assuming I had access to the chatbot in the future, I predict that I would use it	◯	◯	◯	◯	◯
I would use the chatbot frequently if I could	◯	◯	◯	◯	◯

Table C.5: Behavioural Intention to Use

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I preceived the chatbot to be humorous	◯	◯	◯	◯	◯
There was not enough humour in the conversation	◯	◯	◯	◯	◯

Table C.6: Humour

C.3 Open questions

What did you like about the chatbot? _____

What did you not like about the chatbot? _____

(Optional) How can the chatbot be improved? _____

(Optional) Was there anything unusual while you talked to and played with the chatbot? _____

The questionnaire is available through Qualtrics, the link of which can be provided upon request.

C.4 Debriefing Text

Thank you for your participation in this research on the effect of external variables on the user experience of a chatbot. Three conditions were tested against a neutral condition: an emotional chatbot, a chatbot that expresses humour and an aggressive chatbot. The aim of this study is grasping a better understanding of what affects user experience of chatbots. Apologies if the chatbot frustrated you because it played very well.

All results are grouped together; therefore, individual results are not available. Your participation will remain confidential. The final results of this study will be available on the 18th of June. If you would like to read about the research, you can contact us after that date. If you have any remaining questions regarding this research, please contact m.vandelockant@student.ru.nl.

For SurveyCircle-users (www.surveycircle.com):
The Survey Code is: THAS-V6JZ-8F62-M2AN

Appendix D

Connect Four Implementation

The code of the Connect Four website, game logic and chatbot can be provided upon request.