Exploring the Impact of User Interface Design in the Human-Chatbot Interaction

Saralin Zassman
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
szassman@uwaterloo.ca

***Abstract*—This paper investigates how user interface design affects the human-chatbot interaction. We used previous findings on colour, typography, and avatars to produce two chatbots with different levels of visual appeal. The chatbots use a rule-based DialogFlow agent for response generation. We conducted a design exercise where participants interacted with a chatbot, either with a simple design referred to as NEUTRAL or a visually appealing design referred to as HAPPY. After the interaction, participants used a questionnaire to evaluate their emotional state and overall experience with the chatbot. Although differences were not statistically significant, user experience ratings for the HAPPY chatbot were higher than those for the NEUTRAL chatbot. Both groups experienced a similar level of positive affect, suggesting that chatbot design does not have a strong impact on the emotional state of users. However, further investigation is required to support this claim, as well as determine the correlation between affective response and the chatbot user experience.**

***Index Terms*—human-computer interaction, affective computing, chatbot, conversational agent, visual aesthetics, UI design**

I. Introduction

In recent years, there has been a dramatic increase in the popularity of chatbots [1]. Chatbots are now used in a plethora of different domains, such as education, internet support, and customer service. They are also built into smartphones as virtual assistants like Apple's Siri and Microsoft's Cortana. Since they have become such an integral part of the modern world, it is important that we study how different factors influence the human-chatbot interaction. Human-computer interaction (HCI) is a field of study that focuses on understanding how humans and machines interact. Chatbots are software applications that use natural language to interact with users. The motivation for this work stems from the lack of research on the human-chatbot interaction. Ciechanowski *et al.* [2] argue that "little attention has thus far been paid to the socio-cognitive nature of the interaction between man and technology in general and chatbots in particular".

Emotions form the basis of decision making, and are therefore determinants of visual appeal [3]. Affective computing is the study of technology that relates to, impacts, or imitates emotion. Norman [4] argues that emotional design can have a stronger impact on user experience than usability. Emotional design refers to products that are created with the intention to produce positive emotional responses. Norman's design model claims that human affect and behaviour when using a product are the outcome of three levels of processing: the visceral level, the behavioral level, and the reflective level. The visceral level refers to immediate, subconscious responses (i.e., first impressions). At this level, the design components of a system (e.g., symmetry, cleanliness, perceived aesthetics) dictate human perception. Lindgaard *et al.* [5] determined that participants form impressions after being shown a stimulus for only 50 milliseconds. In the context of chatbots, the visceral level refers to involuntary emotional responses to UI design (i.e., colour, font type, and layout). At the behavioural level, consumers form opinions of a product based on its usability and performance, or for our purposes, how well the chatbot responds to the user. The third level of processing is the reflective, where the message and purpose of a product are used to form an an overall impression. Norman's model provides a useful foundation for the design and evaluation of chatbots.

The idea of a chatbot has been credited to Alan Turing who proposed the infamous Turing Test to answer the question, "Can machines think?" [3]. A computer system passes the Turing Test if it can successfully fool people into thinking it's human. Therefore, a chatbot passes the Turing Test when users are unable to distinguish whether they are speaking to a human or a computer. Several chatbots were created to try and pass the Turing Test but none of them have succeeded. ELIZA, the first conversational agent developed in 1966 uses basic NLP techniques to rephrase user input and mimic a psychiatrist [6]. The success of ELIZA lead to a surge in the creation of chatbots, with many hoping that their design would pass the Turing Test. However, chatbots are no longer created solely for entertainment or to mimic human conversation, with many being used for functional purposes such as education, information retrieval, and e-commerce [7]. A growing number of organizations now use chatbots to increase response time for customer support and reduce operational costs [8].

Despite advancements in chatbot technology, to the best of our knowledge, no existing works focus on the UI design of chatbots. We conducted a design exercise to understand the connection between chatbot aesthetics and emotion, as well as their impact on user experience (UX). Participants interacted with a chatbot, either with a simple UI design referred to as NEUTRAL or a visually appealing UI referred

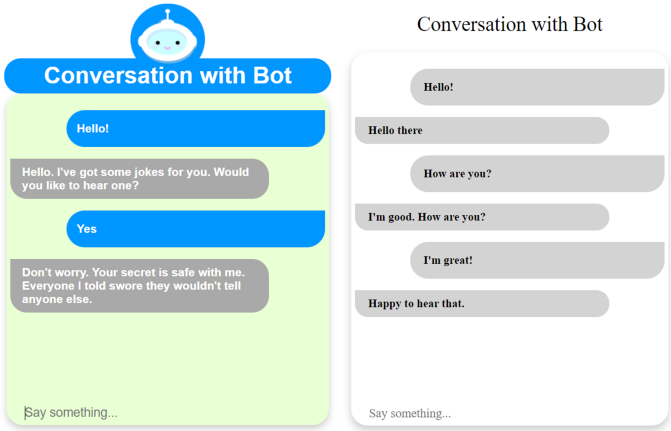

Fig. 1. UI of the HAPPY chatbot (left) and the NEUTRAL chatbot (right)

to as HAPPY. We incorporated previous findings on colour, typography, and avatars in the design of each chatbot. Figure 1 shows the design of each chatbot. Since there may be a strong carryover effect, each participant interacted with only one of the chatbots. We used the User Experience Questionnaire (UEQ) to compare the UX of each chatbot, and the Positive and Negative Affect Schedule (PANAS) to determine the impact of UI design on users' affective responses.

The main hypothesis of the design exercise is that significant differences in UX and positive affect are found in relation to UI design. We predict that positive affective responses will be most intense in the interaction with the HAPPY chatbot but reduced in the interaction with the NEUTRAL chatbot. We also hypothesize that the HAPPY chatbot will induce a better UX than the NEUTRAL chatbot.

II. Response Generation

Task-oriented chatbots are designed to achieve a certain goal, such as booking a flight or ordering food. Non-task-oriented chatbots do not have a particular goal and are used more for entertainment purposes. Often referred to as social chatbots, these systems are designed to hold unstructured, human-like conversations, and were originally created for psychological therapy [9]. We used a non-task-oriented chatbot for the user study.

Response generation forms the core of the chatbot architecture, and takes form either through rule-based or data-driven approaches. Rule-based models are built from a set of predefined rules. These chatbots are easy to create and perform well for simple tasks. ELIZA is a rule-based chatbot that uses pattern and transform rules to generate responses [6]. ELIZA searches a knowledge base to determine the word with the highest rank (i.e., the keyword) in a user expression. The transform rule associated with the keyword is then used to form a response. Rule-based chatbots use a finite set of rules for a practically infinite number of potential messages. This can severely limit the conversational ability of the system and responses may become predictable after a certain point.

Data-driven chatbots make use of large datasets of text, either through information retrieval or generative-based methods. Information retrieval-based systems search a database of candidate responses and match responses using specific sentence structure or machine learning techniques. The key advantage of this approach is that there is complete control over potential responses and thus, the chatbot is unable to send unexpected or inappropriate messages. Modern chatbots like Echo and Siri use results from web searches and advanced information-retrieval techniques to formulate responses [10]. Generative-based models create new responses using vast amounts of training data. Sequence-to-sequence (seq2seq), the current state-of-art model for response generation, uses Recurrent Neural Networks (RNN) to transform user expressions into output messages [9]. Hybrid models exist that generate new responses when retrieval-based techniques fail (i.e., when the input message does not match a candidate response in the database).

III. Related Work

A. Aesthetics and Usability

Several studies have found a connection between visual aesthetics and perceived usability. Kurosu and Kashimura [11] conducted a study using 252 students in Japan to examine the factors that influence perceived usability. They used 26 different ATM layouts and varying levels of usability for the study. The results determined that perceived aesthetics has a greater impact on perceived usability than inherent usability. Tractinsky [12] conducted a follow-up study to determine if the same results could be obtained using 104 students in Israel. The results were consistent with Kuroso and Kashimura's [11] study, suggesting that there is a strong correlation between perceived aesthetics and perceived usability across cultures.

Oyigbo and Vassileva [13] used four mobile website designs to study the relationships between aesthetics, usability, and credibility. They conducted a cross-cultural study with over 500 participants from five different continents. They discovered that visual appeal enhances the perceived ease of use and credibility of mobile websites, and that this effect is irrespective of gender. However, the websites were not actually used in the experiment so results were based on perception.

As a follow-up, Tractinsky *et al.* [14] used nine of the 26 ATM layouts to determine how aesthetics impact implicit and explicit impressions. They discovered that perceived usability after use was influenced by perceived aesthetics and not by the actual usability of the system. The results indicate a strong correlation between beauty and usability, hence the title: "What is beautiful is usable". Hassenzahl [15] challenged this idea in a similar experimental study using MP3 players. He discovered that inherent usability has a greater impact on perceived usability than aesthetics, which contradicts the findings of Tractinsky *et al.*.

A more recent study [16] determined that visual aesthetics only play a major role in perceived usability in the short term. Minge and Thuering used portable audio players for the experiment. The visually appealing audio player had a

symmetric design, a curved body shape, and high colour differences (i.e., blue and grey). The less visually appealing audio player had an asymmetric design, an angular body shape, and low colour differences (i.e., blue and green). After each of three phases, participants assessed their emotional state, as well as the usability and visual aesthetics of the audio player. First impressions were captured in the pre-use phase (i.e., before the system was used). Participants used the audio player for two minutes in the exploration phase, and then performed a number of tasks with the audio player in the final phase. The influence of system usability on perceived usability increased over the phases whereas the influence of visual aesthetics decreased, especially after the pre-use phase. In the context of Norman's design model, this suggests that judgements formed at the behavioural level outweigh those formed at the visceral level.

B. Chatbots

As demonstrated in the previous section, several studies have investigated the aesthetic-usability effect. Here, we discuss the few works that have examined aesthetics in relation to chatbot design.

Meyer-Waarden *et al.* [17] identified a set of factors that influence the intention to reuse an airline customer service chatbot. They determined that colour and visual appeal have a significant positive effect on perceived usability and perceived ease of use. However, they did not include how visual appeal was measured or the specific colours that were used.

Daher *et al.* [18] conducted a between-user study to determine how an avatar impacts the human-chatbot interaction. One version of the agent used text to communicate with the user and express emotions while the other used text and a 2D cartoon avatar with facial and body expressions. They found that the chatbot with the avatar produced more positive affects and less negative affects. However, the difference was not statistically significant and a small sample size was used.

Ciechanowski *et al.* [2] compared the negative affect and uncanny valley effect produced by a simple text chatbot and a chatbot with avatar and sound. The more advanced chatbot used text, speech, and a human-like avatar that moved to match the speech being generated. Participants experienced more negative emotions and a stronger uncanny valley effect when using the chatbot with the avatar. Our design uses a 2D robot avatar to avoid producing an uncanny valley effect.

To the best of our knowledge, no previous works focus on the role of visual aesthetics in the human-chatbot interaction. Meyer-Waarden *et al.* [17] investigated a variety of other factors whereas Ciechanowski *et al.* [2] and Daher *et al.* [18] centred their work on the impact of avatars.

IV. Interface Design

We used the results from the following studies in the design of each chatbot. The HAPPY chatbot was designed to evoke positive affects whereas the NEUTRAL chatbot was designed to evoke neutral affects, or less positive affects than the HAPPY chatbot.

Adams and Osgood [19] conducted a cross-cultural study using participants from 20 countries to determine the connection between colour and affect. The results were generally consistent with the findings of 89 previous studies. Colour and blue were the most highly evaluated concepts, followed by green and white. Black and grey were associated with bad or passive emotions. Hall and Hanna [20] studied the impact of web page text-background colour combinations. They found that grayscale colour combinations (i.e., white, black, and grey) were associated with lower aesthetic ratings.

Tsonos *et al.* [21] measured emotional response in relation to font type and font/background colour combinations. Arial evoked a more pleasant emotional response than Times New Roman. Strong contrast between font colour and background colour was associated with pleasure. Wang and Fodness [22] created two travel retail websites to determine the impact of a 2D avatar on the consumer experience. They found that participants experienced more trust and positive affects towards the website with the avatar than the one without. This, however, was contingent on whether participants found the avatar likeable.

Lee and Koubek [23] determined that colour, layout, and font type can successfully be used to alter the perceived aesthetics of a web page. They used the results of previous studies to produce different levels of visual appeal. The web page with high aesthetics used an analogous colour scheme (orange, yellow, and green), appealing font types, and a layout based on the principles of visual layout, which they define as proximity, alignment, consistency, and contrast. The web page with low aesthetics used grayscale colours, unappealing font type that alternate within the body text, and awkward layouts. They reported a statistical difference between the perceived aesthetics of each web page.

To achieve a high level of aesthetics, the HAPPY chatbot has an analogous colour scheme, an appealing font type (Arial), and a 2D robot avatar. It mainly uses blue and green, which were the highest evaluated colours from Adams and Osgood's study [19]. The NEUTRAL chatbot has a grayscale colour scheme and an unappealing font type (Times New Roman). Both designs have average colour differences between font and background colour, and the same font size and general layout to ensure equal readability.

V. Methodology

A. DialogFlow

Google's DialogFlow[1] is one of the most popular frameworks for chatbot implementation. DialogFlow uses a rule-based approach to generate responses. Intent refers to the end-user's intention with a message. DialogFlow maps every input message to a particular intent. For example, the prebuilt small-talk agent maps "how old are you" and "tell me your age" to an intent built specifically to handle questions about age. An intent includes training phrases and responses. Training phrases are expressions that match what the user might say.

[1] https://cloud.google.com/dialogflow/

End-user expressions must be similar to one of these phrases in order to match the intent. DialogFlow uses machine learning to expand the list of training phrases so only a small number are required. The "Default Fallback Intent" is triggered when an input message does not match any of the other intents, and contains responses like "I do not understand" and "I'm sorry. I didn't quite grasp what you just said".

The chatbots in our design exercise used DialogFlow to make small-talk and tell jokes. We used the prebuilt small-talk agent provided by DialogFlow, with minor adjustments to improve overall quality. To reduce any impact the text may have on emotional state, we added neutral responses and removed overly emotional responses from the agent. Since the small-talk intents only cover a limited number of topics, the chatbot is also able to tell jokes. While a joke-telling chatbot may elicit a more positive response, we make an exception here since both chatbots use the same set of jokes and the jokes are more pun-based than emotionally stimulating.

B. Participants

Friends, family, and classmates were used for the design exercise. The majority of participants were enrolled in or have completed post-secondary education. The mean age of participants was 29.5 with a standard deviation of 16.6. 15 participants used the HAPPY chatbot (M=31.6;SD=20.0 years) and 17 used the NEUTRAL chatbot (M=27.6;SD=13.4 years). The groups are uneven because we used a random redirect tool[2] to assign a chatbot to each participant (i.e., the same link was sent to all participants). The purpose of the design exercise was hidden from participants and only one chatbot was shown to each participant.

C. Questionnaire

After using the chatbot, participants completed the UEQ by Laugwitz *et al* [24] and the positive affect schedule from the PANAS by Watson *et al.* [25].

The UEQ includes 26 items that are each rated on a 7-point Likert scale. Each item uses a pair of opposite adjectives (e.g., annoying/enjoyable, friendly/unfriendly), and the answers range from -3 (strongly agree with the negative term) to 3 (strongly agree with the positive term). The items are divided into six scales: Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. While several existing works focus on the usability of chatbots, we used the UEQ because it measures both pragmatic and hedonic quality. Figure 2 shows the assumed scale structure of the UEQ. Perspicuity, Efficiency, and Dependability are task-oriented, pragmatic quality aspects. Stimulation and Novelty are non-task-oriented, hedonic quality aspects, and Attractiveness should depend on the other five scale values.

The PANAS questionnaire includes 20 items with 10 positive and negative affective descriptors, such as "interested" and "excited". Participants recorded responses for each item using a 5-point Likert scale that ranges from "very slightly or not at all" to "extremely". The test is based on a model that proposes two independent dimensions of emotion. The positive dimension refers to emotional well-being where people feel enthusiastic, alert, or happy. The negative dimension is associated with feelings of anger or sadness (i.e., an emotional state of distress). The sum of all items determines the total score for each dimension, with scores ranging from 10–50.

[2]https://allocate.monster

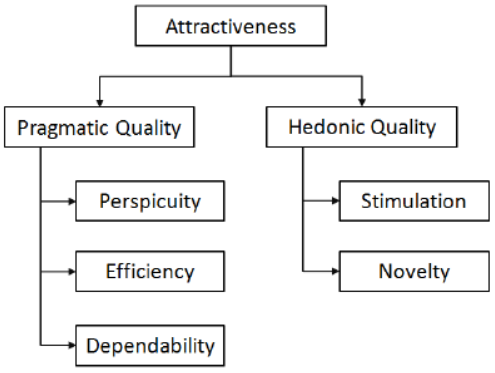

Fig. 2. Assumed scale structure of the UEQ

VI. Results

A. User Experience Questionnaire

The UEQ was used to assess the user experience of each chatbot. The results were evaluated through a two-sample t-test with 95% confidence. There was no statistically significant difference for all six UEQ scales, as demonstrated in Figure 3 by the overlapping confidence intervals. The HAPPY chatbot had a higher mean score than the NEUTRAL chatbot for all six dimensions. The ratings for all six scales of the HAPPY chatbot were positive. The novelty of the NEUTRAL chatbot was the only negatively rated scale. Stimulation had the largest difference in scale mean between the chatbots and perspicuity had the smallest difference in scale mean. The differences between individual items were not statistically significant ($p < 0.05$), and almost all of the ratings had a variance of 1.5 or above. The majority of items in the UEQ were rated higher for the HAPPY chatbot than the NEUTRAL chatbot (i.e., 20 out of 26).

B. Positive Affect Evaluation

There was no statistically significant difference in positive affect between the two groups. Table I summarizes the statistical properties of the positive affect evaluation. For each chatbot, we report mean, standard deviation, median, minimum, and maximum. These measurements demonstrate how similar the results are for each chatbot. The HAPPY chatbot and the NEUTRAL chatbot both had a mean positive affect score of 33, and medians of 34 and 33 respectively. Figure 4 shows the mean values for each item in the positive affect schedule with 95% confidence intervals. None of the differences in individual items are statistically significant, as seen by the overlapping confidence intervals.

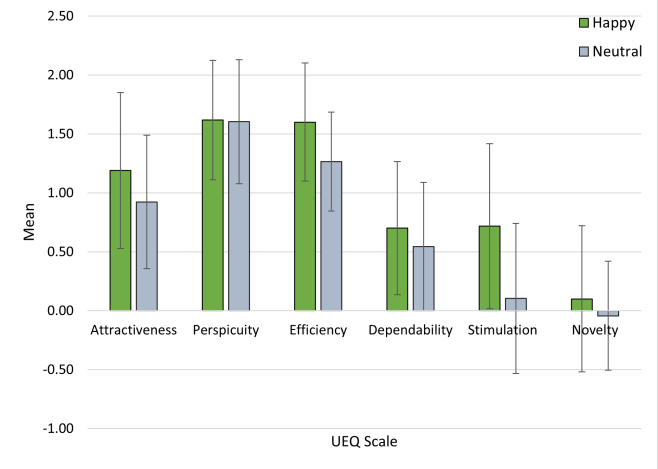

Fig. 3. Results of the UX Evaluation according to the six UEQ scales

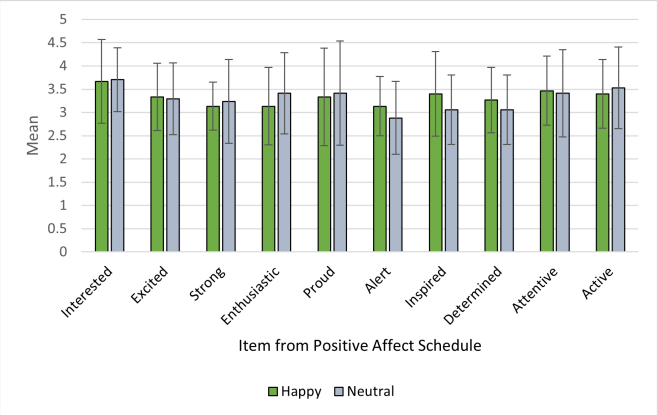

Fig. 4. Results of the Positive Affect Evaluation

VII. Discussion

Since we did not conduct a formal study, the results from the design exercise are used as an estimate on the effects of UI design, and the relationship between positive affect and the chatbot interaction.

A. User Experience

The results do not support our hypothesis stating that there is a statistically significant difference in UX between the chatbots. Although none of the differences were significant, the HAPPY chatbot had a higher average score than the NEUTRAL chatbot for all six scales. Since the two chatbots are identical apart of UI design, this suggests that visual appeal may improve the human-chatbot interaction. Overall, the results indicate several potential claims that require further investigation.

Attractiveness measures the general impression of a product, so participants liked the HAPPY chatbot more than the NEUTRAL one. We expected a significant difference for this scale since the HAPPY chatbot was designed to be more appealing. In terms of pragmatic quality, the HAPPY chatbot was perceived as more efficient and slightly more dependable,

TABLE I
Descriptive statistics of the Positive Affect Evaluation

	Happy Chatbot	**Neutral Chatbot**
Mean	33.27	33.00
SD	5.96	6.34
Median	34	33
Min	21	21
Max	41	43

TABLE II
Selected items of the UEQ

	Happy Chatbot		**Neutral Chatbot**	
Item	Mean	Variance	Mean	Variance
easy/hard to learn	0.7	1.9	-0.1	2.2
easy/complicated	1.9	1.7	2.1	1.4

but equally as perspicuous as the NEUTRAL chatbot. Since previous works [11], [12], [14], [17] found a strong correlation between perceived aesthetics and perceived usability, we believe significant differences in pragmatic quality may emerge with a formal study and slight modifications (which are described in Section VIII). Alternatively, existing works [15], [16] determined that aesthetics only impacts perceived usability in the pre-use phase, and thus, differences may be insignificant because the chatbots were evaluated after use.

Both chatbots had similar ratings for perspicuity (see Figure 3), which suggests that UI design does not have a strong impact on perceived ease of use. This disagrees with the findings of previous works [13], [17] that found a strong connection between aesthetics and perspicuity. Interestingly, the HAPPY chatbot had a higher mean for the easy/hard to learn item and a lower mean for the easy/complicated item, as seen in Table II. This means that participants found the HAPPY chatbot easier to learn and the NEUTRAL chatbot easier to use. These are two somewhat opposing ideas that require further investigation. We presume the visual appeal of the HAPPY chatbot enhanced its perceived ease of use, whereas the simple UI (i.e., no avatar, few colours) of the NEUTRAL chatbot made it appear less complicated.

In terms of hedonic quality, the HAPPY chatbot was seen as more stimulating and novel than the HAPPY chatbot. In fact, the stimulation scale had the largest difference in scale mean between the two groups, as seen in Figure 3. This indicates that participants found the HAPPY chatbot more fun and exciting. As expected, both chatbots had low ratings for novelty and hence, were not seen as particularly innovative or creative. This is likely due to the agent's limited conversational skills, and the fact that modern chatbots are much more complex.

B. Positive Affect

Both groups experienced relatively positive levels of affect after using the chatbot. Although the agent's responses were altered to be more neutral, the joke-telling functionality may

have raised the overall scores. The results do not support our hypothesis that the HAPPY chatbot evokes more positive affects. Even though one of the chatbots was designed to produce more of an emotional response, both groups experienced the same level of positive affect. Therefore, chatbot UI design may not have a significant impact on the emotional state of users. Since both chatbots use the same DialogFlow agent and generate similar responses as a result, conversational style and ability may be more at play, overshadowing any effect that UI design may have. For future work, we plan to study the level at which different components like UI design and response style affect emotional state.

We believe the design of the NEUTRAL chatbot may have caused differences for both evaluations to be statistically similar. Although the NEUTRAL chatbot uses less visually appealing design components like a grayscale colour theme and Times New Roman font, the overall design itself is more simple than visually unappealing. Previous works that found a significant aesthetic-usability effect altered the layout and readability of the less appealing design. We created more of a subtle difference in visual appeal between the designs which may have caused the differences in positive affect and UX to be less pronounced.

From our results, we are unable to determine whether there is a correlation between positive affect and user experience. Since both chatbots produce similar positive affects and user experience ratings, the two may be interrelated (i.e., user experience of a chatbot impacts affective state and vice versa). However, significant differences in both evaluations would need to occur in order to confirm or reject this claim.

VIII. Limitations

The limitations of the design exercise included the use of a small sample size and participant bias. Since around 15 participants used each chatbot and there was a large amount of variation in some of the measurements, different results could have occurred with a larger sample size. In terms of participant bias, some participants may have been more familiar with technology than others. While the majority of participants were in their early 20s, five of the participants were over 50 with three in the control group and two in the experimental. Since older people tend to use less technology, these participants may have had lower expectations and given higher ratings as a result. Participants with a background in computer science or that use virtual assistants like Siri or Alexa on a regular basis may have been more critical. In fact, one of the participants who works in the tech industry (and was unaware of the purpose of the questionnaire) expressed that there were several aspects of the chatbot that could have been improved. Similarly, participants that use less technology are more likely to be less critical and give higher ratings. Despite the clear conversational limitations of the chatbot, some participants reported that they were quite impressed, particularly with the joke-telling functionality. We also realize that friends and family may give higher ratings due to conflict of interest. This would explain the relatively positive ratings for both chatbots despite the major limitations of the system.

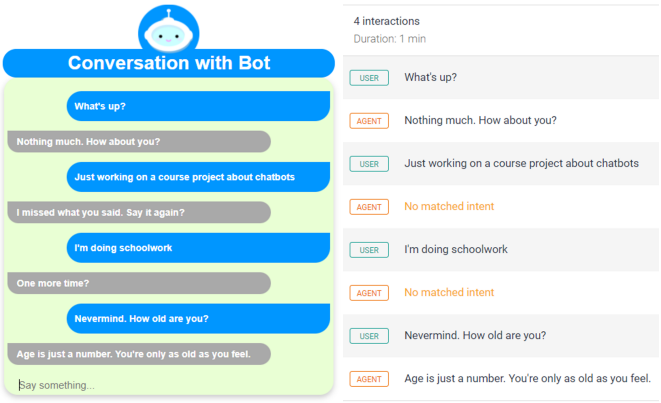

Fig. 5. Example conversation from the HAPPY chatbot (left) and DialogFlow console (right)

For future work, we plan to eliminate the discrepancies described above by recruiting participants from the same age group that have similar backgrounds in technology and clear expectations of how a chatbot should perform. This will prevent any overly positive ratings and potentially highlight the differences between groups. We also plan to conduct a cross-cultural study so our results represent more than a small subset of the human population. This is especially important because chatbots are used all over the world, and the Western industrialized population (i.e., the participants used in our design exercise) has shown significant differences in visual perception when compared to other populations [26]. However, a cross-cultural study may be difficult to conduct given that visual appeal is culture-dependant [27].

The chatbot is very limited in terms of conversational ability. By using a rule-based approach, the DialogFlow agent uses the Fallback Intent whenever an input message does not match an intent—which can be quite often since the existing intents only cover a small set of potential topics. In Figure 5, the chatbot seems to have trouble understanding the user because it has no intent associated with course projects or schoolwork. However, since the chatbot has a specific intent for age-related questions, it is able to give an appropriate response for the last message. Many participants expressed frustrations when the chatbot was unable to tell the date, time, or weather. These are some of the basic and more popular features of virtual assistants. Due to technical limitations, we were unable to add these functionalities to the DialogFlow agent. For future work, we plan to use a more advanced chatbot that compares to other systems on the market. If users are able to interact with the chatbot as they would with a customer service agent or virtual assistant, the affects of UI design may become more apparent.

The issue with evaluating chatbots is there is no way to control how the user interacts with the system. While some users may ask simple questions that the chatbot is able to respond to, others may use more confusing prompts that highlight the agent's weaknesses. We noticed that some users

focused on the joke-telling functionality of the agent while some tried to use the agent as if it were a virtual assistant. These two interactions would yield very different interactions because the agent is unable to answer basic questions that most virtual assistants can. A potential solution is to have participants rate the chatbot after watching a video of someone else using it. However, this may restrict the scope of our findings to the visceral level of perception.

Another limitation is participants used the chatbot for varying amounts of time. As we discussed in Section III, a recent study [16] discovered that the effects of visual appeal vanish after a product is used for a sufficient amount of time. As a result, reviews that were more first-impression-based were likely more affected by UI design. When we conduct a formal study, we plan to ensure each participant interacts with the system for the same amount of time.

IX. Conclusion

In this design exercise, we investigated the impact of UI design on the human-chatbot interaction. We used previous research on aesthetics and emotion to create two chatbots with different levels of visual appeal. With 32 participants, we compared the performance of each chatbot using the UEQ and the positive affect schedule from the PANAS. The differences between chatbots for both evaluations were statistically insignificant, and thus, the results indicate that visual appeal does not impact the UX of chatbots. However, the visually appealing chatbot had a higher rating for all of the scales in the UEQ. This suggests that significant differences in UX may occur with a larger and less biased sample. For the positive affect evaluation, the results were very similar for both groups, even though the HAPPY chatbot was designed to evoke more positive emotions. Overall, we believe further investigation is required to determine how visual appeal impacts the chatbot interaction.

X. Future Work

A clear direction for future work would be conducting a formal study with a larger, more diverse sample, as well as using a more sophisticated agent to generate responses. We also plan to measure positive and negative affect since we predict participants may be more easily irritated when using the less appealing design. It may be interesting to use a task-oriented and non-task-oriented chatbot to determine if one type is more affected by aesthetics than the other. Since questionnaires only capture self-reported emotions, another direction would be measuring both declarative and psychophysiological responses, which was done by Ciechanowski *et al.* [2]. Finally, we could also have participants rate a set of designs prior to the experiment, and then use the design with the highest rating for the HAPPY chatbot.

We plan to conduct an extension of our design exercise that compares UI design with communication style. This would use one chatbot with engaging or human-like responses and another with a more visually appealing UI and neutral responses. Both chatbots would be compared to a baseline system that uses a simple UI and neutral text. A similar study could also be conducted for usability (i.e., visually appealing chatbot with poor usability and vice versa). We are also interested in studying how the impact of chatbot UI design varies with time or number of interactions.

In our design exercise, we used colour, font-style, and an avatar to produce a visually appealing UI. By combining several design components, we are unable to understand the extent to which each component elicits positive affect. For future work, we plan to conduct a separate experiment for each component (e.g., colour, font-style, avatar/no avatar).

Another direction would be to design multiple UIs that each attempt to evoke a distinct emotion, such as fear, happiness, or anger. This was our original idea but we decided to focus on positive affect and leave other affects for future work. This study may be more difficult to implement because there are less visual cues associated with emotions like fear or surprise. The results also have less applications because products are typically designed to make people happy.

XI. Appendix

- HAPPY chatbot: https://chatbot886h.herokuapp.com/
- NEUTRAL chatbot: https://chatbot886n.herokuapp.com/

References

[1] L. C. Klopfenstein, S. Delpriori, S. Malatini, and A. Bogliolo, "The rise of bots: A survey of conversational interfaces, patterns, and paradigms," in *Designing Interactive Systems*, ACM, 2017, pp. 555–565. DOI: 10.1145/3064663.3064672.

[2] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, "In the shades of the uncanny valley: An experimental study of human-chatbot interaction," *Future Generation Computer Systems*, vol. 92, pp. 539–548, Feb. 2018. DOI: 10.1016/j.future.2018.01.055.

[3] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997, ISBN: 978-0-262-16170-1.

[4] D. Norman, *Emotional design : why we love (or hate) everyday things*. Basic Books, 2004.

[5] G. Lindgaard, G. Fernandes, C. Dudek, and J. Brown, "Attention web designers: You have 50 milliseconds to make a good first impression!" *Behaviour & IT*, vol. 25, pp. 115–126, 2006. DOI: 10.1080/01449290500330448.

[6] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966. DOI: 10.1145/365153.365168.

[7] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in *Artificial Intelligence Applications and Innovations*, 2020, pp. 373–383.

[8] E. Almansor and F. Hussain, "Survey on intelligent chatbots: State-of-the-art and future research directions," in *Complex, Intelligent, and Software Intensive Systems*, 2020, pp. 534–543.

[9] M. Mnasri, "Recent advances in conversational nlp : Towards the standardization of chatbot building," *ArXiv preprint arXiv:1903.09025*, 2019.

[10] J. Cahn, "Chatbot: Architecture, design, & development," Senior thesis, University of Pennsylvania, 2017.

[11] M. Kurosu and K. Kashimura, "Apparent usability vs. inherent usability experimental analysis on the determinants of the apparent usability," in *CHI*, vol. 2, Jan. 1995, pp. 292–293. DOI: 10.1145/223355.223680.

[12] N. Tractinsky, "Aesthetics and apparent usability: Empirically assessing cultural and methodological issues," in *CHI*, 1997, pp. 115–122. DOI: 10.1145/258549.258626.

[13] K. Oyibo and J. Vassileva, "The interplay of aesthetics, usability and credibility in mobile websites and the moderation by culture," in *IHC: Brazilian Symposium on Human Factors in Computing Systems*, ACM, Oct. 2016, pp. 1–10. DOI: 10.1145/3033701.3033711.

[14] N. Tractinsky, A. Katz, and D. Ikar, "What is beautiful is usable," *Interacting with Computers*, vol. 13, pp. 127–145, 2000. DOI: 10.1016/S0953-5438(00)00031-X.

[15] M. Hassenzahl, "The interplay of beauty, goodness, and usability in interactive products," *Human-Computer Interaction*, vol. 19, pp. 319–349, Dec. 2004. DOI: 10.1207/s15327051hci1904_2.

[16] M. Minge and M. Thuering, "Hedonic and pragmatic halo effects at early stages of user experience," *International Journal of Human-Computer Studies*, vol. 109, 2017. DOI: 10.1016/j.ijhcs.2017.07.007.

[17] L. Meyer-Waarden, G. Pavone, T. Poocharoentou, *et al.*, "How service quality influences customer acceptance and usage of chatbots?" *Journal of Service Management Research*, vol. 4, pp. 35–51, 2020. DOI: 10.15358/2511-8676-2020-1-35.

[18] K. Daher, Z. Bardelli, J. Casas, E. Mugellini, O. A. Khaled, and D. Lalanne, "Embodied conversational agent for emotional recognition training," Advances in Computer-Human Interactions, 2020, pp. 384–390, ISBN: 9781612087610.

[19] F. M. Adams and C. E. Osgood, "A cross-cultural study of the affective meanings of color," *J. Cross-Cult. Psychol*, vol. 4, no. 2, pp. 135–156, 1973. DOI: 10.1177/002202217300400201.

[20] R. Hall and P. Hanna, "The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention," *Behaviour & IT*, vol. 23, pp. 183–195, May 2004. DOI: 10.1080/01449290410001669932.

[21] D. Tsonos, G. Kouroupetroglou, and D. Deligiorgi, "Regression modeling of reader's emotions induced by font based text signals," *Lecture Notes in Computer Science*, vol. 8010, pp. 434–443, Jul. 2013. DOI: 10.1007/978-3-642-39191-0_48.

[22] L. C. Wang and D. Fodness, "Can avatars enhance consumer trust and emotion in online retail sales?" *IJEMR*, vol. 3, no. 4, pp. 341–362, 2010. DOI: 10.1504/IJEMR.2010.036881.

[23] S. Lee and R. Koubek, "The impact of cognitive style on user preference based on usability and aesthetics for computer-based systems," *International Journal of Human–Computer Interaction*, vol. 27, pp. 1083–1114, Nov. 2011. DOI: 10.1080/10447318.2011.555320.

[24] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *HCI and Usability for Education and Work*. Nov. 2008, vol. 5298, pp. 63–76, ISBN: 978-3-540-89349-3. DOI: 10.1007/978-3-540-89350-9_6.

[25] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The panas scales," *Journal of personality and social psychology*, vol. 54, no. 6, pp. 1063–1070, 1988. DOI: 10.1037//0022-3514.54.6.1063.

[26] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?" *Behavioral and Brain Sciences*, vol. 33, no. 2-3, pp. 61–83, 2010. DOI: 10.1017/S0140525X0999152X.

[27] K. Reinecke and A. Bernstein, "Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces," *ACM Transactions on Computer-Human Interaction*, vol. 18, no. 2, Jul. 2011. DOI: 10.1145/1970378.1970382.