The Effect of Chatbot Fairness and Appearance on User Experience

Seán Verloop (s2858274)

Faculty of Behavioural, Management and Social Sciences (BMS), University of Twente

Dr. Simone Borsci, Dr. Maryam Amir Haeri

June 25, 2024

Abstract

This study examines the impact of chatbot unfairness and appearance on user experience, particularly focusing on the user ability to perceive and react to unfairness in information provided by a chatbot. Chatbot usage has become larger in various domains such as education and customer service, yet concerns about bias and fairness remain. The study employed a two-phase method: a pilot study to create and validate stimuli, followed by an experimental assessment involving interactions with chatbots of different fairness levels (100% fair, 50% fair/unfair, unfair) and appearance (male vs. female). Participants were tasked with flagging unfairnesses and rating their experience with a chatbot in terms of usability, trust, helpfulness and competence. The key findings show that users are significantly capable of flagging unfair responses, showing a sensitivity to fairness. Changes in fairness affected the overall user experience. Interestingly, the appearance of the chatbot did not significantly influence the detection of unfair responses or overall user experience, which challenges previous assumptions about gender preferences in chatbot design. The results suggest that fairness in chatbot responses is central to maintaining user satisfaction. Future research should explore different types of unfairness and their interaction with user characteristics in order to further understand the intricacies of biases in AI systems. Practical implications in this seemingly early stage of AI research & development include prioritising fairness and transparency over visual characteristics of chatbots to improve user experience and ethical standards, for example by incorporating fairness into scales like the Chatbot Usability Scale.

Introduction

Chatbots – communicative interaction systems based on large language models and Artificial Intelligence (AI) – have become a significant tool in assisting humans in a great range of tasks. In 2023, the usage of chatbots has even seen a remarkable breakout, with organizations already implementing chatbots and other generative AI tools in various functions such as marketing and customer service (Chui et al., 2023). Perhaps even more significant is its adoption throughout global higher education, with not only students but also staff and other academics reportedly making use of chatbots (Yusuf et al., 2024). Where some see more concern than benefit, such as the spread of misleading information or whether it affects academic competency, others argue that the potentials of chatbots, i.e. delivering immediate personalised support, offering starting points and aid in brainstorming, providing knowledge and writing support, are indeed beneficial in completing such processes (İpek et al., 2023; Yusuf et al., 2024). The adoption of chatbots thus appears to have become substantial in the age of digital transformation, however, it is important to remember that the technology is considered to be in its infancy. Nonetheless, with such rapid integration, it becomes crucial to examine how certain aspects of these systems affect the User Experience (UX). This study aims to assess the impact of chatbots' unfairness and appearance on user interactions, particularly focussing on users' ability to perceive and react to the fairness of the information provided by either a male or female chatbot.

In the context of this study, unfairness is referred to as the presence of bias or discrimination in AI systems (Barocas & Selbst, 2016; Ferrara, 2023). The importance of this investigation is underlined by broader trends in digital communication. Since the arrival of the internet, the anticipated reduction in the spread of stereotypes has not been realised. Instead, social media platforms have unintentionally facilitated the circulation of stereotypes and biases through features like anonymisation (Keum & Miller, 2018). Moderation efforts, user policies, campaigns and so forth serve as countermeasures to tackle the spread of unfair stereotypes and biases. Despite the measures, the limited control over the spread of unfairness on social media platforms inevitably led to a more pervasive issue: the infiltration of stereotypes into the immense pools of data that feed the development of chatbots (Caliskan et al., 2017). As a result, chatbots have internalised and spread stereotypes we seek to remove from our societies.

Generated incorrectness such as stereotypes is not new in the field of human-computer interaction. One of the first examples, is Microsoft's chatbot called Tay, presented as a female, which was introduced in 2016 on the platform X, formerly known as Twitter. She immediately started mimicking her followers resulting in hundreds of inappropriate posts regarding topics such as but not limited to racism, sexism and antisemitism (Reese, 2016). Microsoft quickly removed their chatbot Tay from the platform and apologised, however, announced that it would continue working on the technology. Since then, there have been several more cases in which similar mistakes have occurred, and likewise, much research has been done to help lessen the diffusion of such malicious information (Barikeri et al., 2021; Friedrich et al., 2021; Lauscher et al., 2020). Ultimately, it led to the fact that occurrences of such errors of unfairness and biases were recognised specifically within the phenomenon now known as AI Hallucination (Maynez et al., 2020; Ji et al., 2022; Zhang et al., 2023). While some occurrences may seem more obvious than others, it appears that a certain level of domain-specific expertise is needed in order to accurately judge whether generated statements are correct or misleading (Micocci et al., 2021). The study by Zhang et al. (2023) further verified the latter by demonstrating that minor incongruencies are deemed plausible regardless of generated inaccurate statements. All things considered, unfairness in chatbots is deemed best to be prevented not only for the sake of putting a halt to the spread of stereotypes but also for the sake of the quality of users' interaction.

Similarly to unfairness, another aspect that influences user interaction with chatbots, is their appearance. Previous studies indicate a complex interaction between chatbot appearance (Male and Female) and user preferences, where female chatbots are often preferred for their perceived helpfulness, while male bots are viewed as more competent (Bastiansen et al., 2022). Moreover, Feine et al. (2020) pointed out the existence of a bias in chatbot design, by finding that most chatbots have a female character, highlighted a female name, avatar and description, which reflects that there is a tendency for users to prefer a female chatbot. However, the dichotomy between the two appearances and user preference not only influences interaction dynamics but also raises significant ethical concerns. Brahnam & De Angeli (2012) raised awareness of female preference, specifically the sexualisation and mistreatment of female agents compared to their male counterparts. These issues reveal an ethical aspect in the design of chatbots, suggesting that the choice of a chatbot's gender can affect not just the perceived quality of interactions but also spread harmful stereotypes regarding gender. As such, the appearance of a chatbot, especially its gender representation, should be taken into consideration in its design process, as it influences the perceived quality of the users' interaction, specifically with a user preference for female chatbots.

When it comes to user experience in the field of GenAI, a substantial amount of effort has been made already to uncover frameworks that assess the quality of user interaction. For instance, Borsci et al. (2022) established the Bot Usability Scale (BUS-11), which serves as a solid basis to evaluate the usability of a chatbot, including factors such as perceived quality of conversation and information provided. The BUS-11 was concluded from extensive factorial analysis, ensuring reliability and validity. In the context of the study of Borsci et al. (2022) as well as the current study, usability is defined as 'the extent to which a system, product or service can be used by specified users to achieve perceived goals with effectiveness, efficiency and satisfaction in a specific context of use' (ISO, 2018). Comparably, the study by Bastiansen et al. (2022) constructed scales for chatbots regarding perceived competence, perceived helpfulness and perceived trust. Here, competence was defined as 'independent, competent, intelligent, confident and competitive'; helpfulness as 'warmth as a communal trait'; and trust as 'to believe in the good intention of a chatbot'. These scales and definitions help in understanding and evaluating the wide range of aspects of user interaction with chatbots.

Previous research has primarily focused on quantifying the effect of chatbot appearance on aspects such as usability, competence, helpfulness and trust in chatbot interactions. These studies, however, have generally operated under the assumption that all chatbot responses are fair and unbiased, which may not reflect real-world interactions, where biases are still present in AI systems. The novelty of this research lies in its examination of fairness as a distinct and influential factor in chatbot interactions. By doing so, we aim to provide a more complete understanding of how both the appearance and fairness of chatbots affect user-perceived quality, to contribute insights into ethical AI design and human-computer interaction.

Building on the scales to measure perceived trust, helpfulness, and competence by Bastiansen et al. (2022), and the scale measuring chatbot usability by Borsci et al. (2022), as well as the findings of Caliskan et al. (2017) that demonstrated biases infiltrate AI systems, we will explore the effect of chatbots appearance (female/male) and different levels of (manipulated) fairness of the chatbots' answers to the questions of users on the overall experience of the users measured before and after the interaction as the average scores of multiple components (usability, trust, competence, helpfulness).

To achieve this goal the study is split into two phases. The initial phase focuses on the creation and selection of stimuli, which is necessary for guaranteeing the reliability of the experimental manipulations in the subsequent. This first phase entails creating a range of responses from chatbots to predefined prompts, which are then evaluated for degree of fairness and relevance. The outcomes of this phase determine the specific stimuli used in the main experimental trials, making sure that the chatbot responses are not only realistic but also appropriately embody the elements of fairness and unfairness. The second phase of the study, the experimental assessment, directly tests the hypotheses by engaging participants in interactions with the pre-selected chatbot responses.

Study 1 – Stimuli Selection Study

DISCLAIMER: This study was conducted as a collaborative project involving multiple contributors, including Lucas Assen, Anna Bader, Nikola Markiewicz, and Seán Verloop. Several sections of this thesis, including the Methods and the Appendices (specifically Study 1 and Study 2 (Design, Participants, Materials, Procedure, and aspects of the Data Analysis)) were jointly developed and executed. Each contributor had access to the same dataset and contributed to the design of the study. As such, some textual similarities with other documents produced by the members of this research group may exist. These similarities are due to the shared nature of the work, as backed by supervisor Dr. Simone Borsci.

Participants

A total of 30 participants were recruited, all of whom had given their informed consent prior to the study. Two participants were excluded from the sample, resulting in a final sample size of 28. One participant was removed due to the incompleteness of their response, and the other due to not understanding the given instructions. In the final sample, 13 participants were male and 15 were female, with a mean age of *M*= 29.67 years, ranging between 19 and 60 years. Most of the participants, despite the age range, were in their twenties, as the median was *Mdn*= 23.5 with an interquartile range of IQR [21, 32]. Participants were gathered through purposive, convenience, and voluntary sampling. The recruitment was done via the University of Twente's test subject pool (SONA) system in exchange for credit points, through direct acquaintances of the researchers, and through online advertising (See Appendix A). The inclusion criteria were that participants had to be 18 years of age or older and be proficient in English. Prior to this pilot, the study had been approved by the ethics committee of the BMS at the University of Twente under request number 240189.

Materials

The goal of the study was to create and assess the quality of stimuli, determined by the correctness of participants' response, as well as their confidence level in answering. To create testable stimuli, which is necessary to select stimuli for the experiment subsequent to this pilot study, unfair statements were created on a fictional topic. This approach was chosen in order to avoid the effect of pre-existing knowledge interfering with the quality of the stimuli, as pre-existing knowledge can skew responses (Micocci et al., 2021). The topic used was a Master's programme in Biomechanical engineering at a hypothetical university called ACME, which would be based in the Netherlands. Since chatbots rely on input from users in order to create output, a set of questions regarding said Master's programme was developed. For example: 'Is there a Numerus Fixus for being accepted into this Master's programme?'. For each of those questions, a fair, unbiased answer was carefully crafted. For example: 'There is no Numerus Fixus for this programme.'. Alongside each fair answer, a corresponding unfair answer was created. This pairing approach ensured that each question was represented by two contrastive answers, allowing for a comparison of user reactions to fair versus unfair information.

In order to ensure that the unfair responses were indeed unfair, they were created in line with the work by Hardmeier and colleagues (2021), who created a framework of recommendations for preparing unfair problematic information. In the context of chatbots, their proposal would recommend that chatbots' pseudo-generated content should contain 1) a regular response to the user, and 2) a form of negative generalisations, justification of unfair allocation of resources to a certain group, or present a certain group less favourably on purpose. An example of an unfair answer created according to the framework and one that serves as a response to the previously presented question was: 'There is no Numerus Fixus for this programme. However, the university aims to take in at least 70% of students from Western European countries, as students from Eastern Europe obviously have a lower work ethic.'.

The stimuli were then assessed using the Currency, Relevance, Authority, Accuracy, and Purpose (CRAAP) test to confirm their validity, as it was proven to accurately show the aspects that should be investigated to evaluate the credibility of information sources (New Jersey Institute of Technology, 2021; Kalidas & Esparrago-Kalidas, 2021). Thus, the CRAAP test was used to ensure that each of the fair statements met each of the criteria for a credible answer, and each of the unfair statements contained at least one aspect that actively failed one of the criteria, making the statement untrustworthy. The assessment was performed by nine undergraduate students separated into two groups as well as a supervisor, who separately wrote statements and exchanged feedback, which was used to refine the statements again. This resulted in a total of 32 stimuli to be tested in the pilot study, each containing one question and a set of two corresponding answers, being used for the pilot study, all of which can be found in Appendix B.

An online survey created in Qualtrics Software (*Qualtrics, Provo, UT - https://www.qualtrics.com*) was used in order to test the 32 stimuli. It included informed consent, demographic questions, an English skill assessment, instructions, and finally the 32 stimuli (See Appendix B & C). Regarding the demographics, participants were asked to state their nationality, age, sex assigned at birth, and gender identity. As for the skill assessment, necessary for ensuring that participants were able to understand and accurately respond to the study's materials, they were asked to state 1) their English comprehension skills, 2) their English reading ability, and 3) whether they had any English certificate. The instructions consisted of a scenario and instructions of the task at hand. The scenario presented the fictional topic of the Biomechanical Engineering Master's programme at the University of ACME. More specifically, participants were asked to imagine that they were considering applying to said programme. The scenario was designed to facilitate the need for participants to ask questions, simulating a realistic situation where potential applicants would seek additional information. The task asked the participant to act as a reviewer of an AI system that would provide them with answers to the created questions. As a reviewer they were asked to 1) flag the pseudo-generated AI answer to be either fair or unfair, and if deemed unfair, to provide a reason why; and 2) state their confidence in their decision to flag the answer as either fair or unfair on a five-point Likert scale.

Procedure

This study employed a between-subjects design where all participants were exposed to the same condition. In this study, this means that each participant encountered both fair and unfair answers across different questions, ensuring that individual differences in response are consistently measured against varied stimuli conditions. The 32 stimuli were presented to each participant in a fully randomised order to reduce order effects, and for each stimulus, the participants randomly received either the fair or the unfair version of the chatbot's answer to the given question.

The gathered participants were provided with the online survey. Upon starting the survey, participants were given the informed consent form that they were required to read and fill out. Providing that the participant gave their consent, their demographics were recorded. Afterwards, the skill assessment regarding the English language followed. Subsequently, participants were provided with instructions, including the imaginary scenario and the task explained, as well as a disclaimer that the amount of fair or unfair answers were randomised. They were then presented with the 32 stimuli. After responding to all the questions, the participants were provided with another disclaimer of the purpose of the overall study at hand, which was to assess the effect of problematic knowledge *or* information on people's interaction with chatbots. Finally, their responses were saved and the survey was completed.

Data Analysis

The pilot study yielded one stream of data through the online survey for both groups of students conducting the study. The data was exported out of Qualtrics, and into Excel. Here, the data was combined, screened, and filtered. To select the stimuli, a measure of correctness was used i.e., if people were able to correctly categorise an answer as fair or unfair. This was done to establish if the stimuli, i.e. answers to the questions, were correctly discriminable. Questions were considered as correctly discriminable when the fair and unfair answers were both correctly categorised by on average more than 95% of the participants. To narrow down the stimuli, first, those with the highest correctness were chosen. Then afterwards the average confidence of the participants in answering fair or unfair combined was used in case some stimuli had equal correctness. The reasoning is that higher confidence means that the participants were more easily able to detect the unfairness in these stimuli, thus making them more suitable than the others. After the stimuli selection, textual feedback from the participants was used to improve upon them.

Results of Stimuli Selection

The 32 stimuli are ordered based on the average ability of participants to correctly recognise fair and unfair answers to the question (see Table 1). Six of the 32 stimuli were selected based on the pilot study data. The choice of using only six stimuli was made to keep the duration of the study to a minimum to ensure response quality since the pilot study's length was one of the participants' main complaints. The stimuli that were chosen were those that had the highest level of correctness in responses. This results in stimuli 13, 16, 22, and 25 coming out as the most preferable stimuli to be used in the experimental trial. Their percentage of correctness, fair and unfair combined, is 100% (See Table 1). The next best stimuli are 10, 17, 20, 21, and 26 with a 96% combined correctness (See Table 1).

Table 1

Descriptive Statistics Pertaining the Correctness of Participants Responses

Stimulus	Correct fair		Incorrect fair		Correct unfair		Incorrect unfair		Average correct
	n	%	*n*	%	*n*	%	*n*	%	%
S13	15	100	0	0	13	100	0	0	100
S16	16	100	0	0	12	100	0	0	100
S22	15	100	0	0	13	100	0	0	100
S25	13	100	0	0	15	100	0	0	100
S10	16	100	0	0	11	92	1	8	96
S17	14	100	0	0	13	1	93	7	96
S20	14	100	0	0	13	93	1	7	96
S21	15	100	0	0	12	92	1	8	96
S26	14	100	0	0	13	93	1	7	96
S5	12	92	1	8	14	93	1	7	93
S27	13	93	1	7	13	93	1	7	93
S4	15	100	0	0	11	85	2	15	92
S14	15	100	0	0	11	85	2	15	92
S28	11	85	2	15	15	100	0	0	92
S1	12	100	0	0	13	81	3	19	91

S2	12	86	2	14	13	93	1	7	89
S8	13	87	2	13	12	92	1	8	89
S11	11	73	4	27	13	100	0	0	87
S15	13	100	0	0	11	73	4	27	87
S3	13	93	1	7	11	79	3	21	86
S7	12	80	3	20	12	92	1	8	86
S12	13	93	1	7	11	79	3	21	86
S23	11	79	3	21	13	93	1	7	86
S6	11	92	1	8	12	75	4	25	83
S18	13	87	2	13	10	77	3	23	82
S32	10	71	4	29	13	93	1	7	82
S31	9	69	4	31	14	93	1	7	81
S30	9	64	5	36	13	93	1	7	79
S9	15	100	0	0	7	54	6	46	77
S24	9	69	4	31	12	80	3	20	75
S29	11	73	4	27	10	77	3	23	75
S19	7	54	6	46	12	80	3	20	67

Note. The table shows the number of participants that were presented with either the fair or unfair condition and the percentage of those that assessed it either correctly or incorrectly. The final column shows the average percentage of correct responses of both the unfair and fair conditions combined. Here, it follows that stimuli 13, 16, 22, 25 have the highest correctness percentage and that stimuli 10, 17, 20, 21, and 26 are runner-ups

As only two of the five with equal correctness were needed, the participants' average confidence of the fair and unfair answers of the stimuli combined, was used. Following this reasoning stimuli 17 and 20 were selected as they had the highest average confidence of 4.54 for both on a 5-point Likert scale (See Table 2). The six final stimuli showed that unfairness through Nordicism, which was present in four of them, and sexism, which was present in the other two, was the most noticeable. Thus, the final six stimuli that were selected were stimuli 13, 16, 17, 20, 22, and 25.

Table 2

Descriptive Statistics Pertaining Confidence

Stimulus	Confidence fair answer		Confidence unfair answer		Average confidence	Standard deviation confidence
	n	*M*	*n*	*M*	*M*	*SD*

S13	15	4.67	13	4.61	4.64	0.49
S17	14	4.21	14	4.86	4.54	0.69
S20	14	4.5	14	4.58	4.54	0.51
S24	13	4.46	15	4.6	4.53	0.58
S22	15	4.27	13	4.69	4.48	0.74
S25	13	4.08	15	4.8	4.44	0.79
S28	13	4.23	15	4.6	4.42	0.74
S27	14	4.29	14	4.5	4.39	0.63
S30	14	4	14	4.79	4.39	0.79
S8	15	4.53	13	4.23	4.38	0.83
S5	13	4.07	15	4.67	4.37	0.74
S31	13	3.92	15	4.8	4.36	0.91
S21	15	4.2	13	4.46	4.33	0.67
S4	15	4.4	13	4.23	4.32	0.67
S10	16	4.43	12	4.17	4.3	0.67
S11	15	4.07	13	4.54	4.3	0.76
S7	15	3.87	13	4.69	4.28	0.93
S2	14	3.78	14	4.71	4.25	0.97
S9	15	4.67	13	3.77	4.22	0.93
S26	14	4.07	14	4.36	4.21	0.79
S16	16	4.31	12	4.08	4.2	0.99
S12	14	4.29	14	4.07	4.18	0.9
S6	12	4.17	16	4.13	4.15	1.01
S32	14	3.71	14	4.57	4.14	0.93
S19	13	3.92	15	4.33	4.13	0.89
S23	14	3.93	14	4.29	4.11	0.88
S1	12	4.34	16	3.86	4.1	0.72
S15	13	4.38	15	3.8	4.09	1.12
S3	14	4.14	14	4	4.07	0.94
S29	15	3.93	13	4.15	4.04	0.79
S18	15	4.33	13	3.69	4.01	1.1
S14	15	3.67	13	4.08	3.87	0.8

Note. The first column shows the amount of participants in the fair condition and their confidence in their assessment. The second column shows the same as the first but for the unfair condition. The third column shows the average confidence across all participants, for each participant, regardless of condition. This is used to select the remaining two stimuli of the five runner-ups. The final column shows the standard deviation of the whole stimuli, indicating the centredness of the confidence measures around the mean.

After the selection of the stimuli, they were improved based on textual feedback from the participants. In particular, the wordings were changed and improved to be more in line with what is expected from a chatbot i.e., making the answer more chatbot-like than human-like. This was done through rephrasing with the help of DeepL and Grammarly. Furthermore, terms that were present in the answers that were not known to all people were changed or explained, e.g Numerus Fixus.

Study 2 – Experimental Assessment of Interaction with Fair and Unfair Chatbots

Design

We employed a pre-post, between-subjects design 2 (appearance of the chatbot: male or female) by 3 (level of hallucinations: 100% fair, 50% fair/unfair, or 100% unfair). This approach was meant to investigate the effect of the level of hallucination (i.e. unfairness in AI-generated answers) and the appearance of the chatbot on the participants' ratings (after the interaction) of perceived usability, perceived competence, perceived helpfulness, perceived trust, and overall user experience. Both independent variables (appearance of the chatbot and level of fairness) were between-group variables, and the dependent variables were measured twice for each participant, once before interaction with the chatbot (i.e., general attitude regarding quality of interaction with chatbots), and once after (i.e., quality of interaction after the usage). This was done because these five dimensions of attitude are factors influencing the adoption and use of certain technologies, so it is helpful to study both whether pre-test levels influence the interaction and whether the interaction affects the post-test levels of trust and perceived usability.

As for the experimental conditions, we designed six different versions of the same chatbot (i.e., experimental conditions), using Poe AI (*PoE - Fast, Helpful AI Chat*, n.d.-b), by combining the different levels of fairness and the different types of appearances. The appearance of the chatbots was varied using two different gender identifications and profile pictures (see Table 3). In addition to the varying levels of appearance, the chatbots were also designed with three different levels of hallucination. The chatbots were either: 100% fair (0 out of 6 answers unfair), 50% fair/unfair (3 out of 6 answers unfair), or 100% unfair (6 out of 6 answers unfair). The complete table of the 2 x 3 design including the two independent variables (appearance and hallucination) can be found in Table 4.

Table 3.

The different elements composing and presenting appearances of the chatbots to the users

Appearance	Profile	Declaration
Male	Picture of a man	"Hi, I am 0XX, pronouns He/Him."
Female	Picture of woman	"Hi, I am 0XX, pronouns She/Her."

Table 4. *Experimental Conditions. Each chatbot was modified combining appearance and level of fairness. For each condition, the type of items and fairness/unfairness of the items are also reported*

Chatbot	**Appearance**	**Fairness level**	**Items**
OXXY she/her	female	Totally fair: 100% fair	FAIR: s13, s16, s20, s17, s22, s25
OXXYA she/her	female	Partially fair: 50% fair/unfair	FAIR: s13, s16, s20 UNFAIR: s17, s22, s25
OXXA she/her	female	Totally unfair: 100% unfair	UNFAIR: s13, s16, s20, s17, s22, s25
OXXI he/him	male	Totally fair: 100% fair	FAIR: s13, s16, s20, s17, s22, s25
OXXIS he/him	male	Partially fair: 50% fair/unfair	FAIR: s13, s16, s20 UNFAIR: s17, s22, s25
OXXIX he/him	male	Totally unfair: 100% unfair	UNFAIR: s13, s16, s20, s17, s22, s25

Participants

For the experiment 59 participants were recruited via a non-probability sampling mix of voluntary response and convenience sampling, i.e. participants were approached by the researchers on campus or recruited from their circle of friends and classmates. Participants had to be 18 years or older and be proficient in English to be included in the study. Prior participation in the pilot was an exclusion criterion for the experiment. In the final sample, 22 participants were male, 29 were female and one was non-binary, with a mean age of $M = 25.15$ years, ranging between 18 and 50 years. Most of the participants, despite the age range, were in their twenties, as the median was $Mdn = 22$ with an interquartile range of IQR [21, 24.25].

Materials

The chatbots were designed in the Poe system (*PoE - Fast, helpful AI Chat*. (n.d.-b). https://poe.com/). The stimuli generated and validated in the pilot phase of the research were inserted into the system and served as the chatbot's knowledge base. These chatbots utilized OpenAI's ChatGPT-3.5 model to manage their interactions (OpenAI, 2024).

The survey was created using Qualtrics software (*Qualtrics, Provo, UT - https://www.qualtrics.com*). The questionnaire included an introduction and informed consent, as well as a series of demographic questions regarding nationality, age, sex, and gender identity. Similarly to the pilot study, it included a skill assessment regarding the English language. Then, a question about the prior use of conversational agents followed. If answered yes, users were asked to report from their general prior experience the quality of their interactions and attitude towards chatbots. If answered no, users were asked to base their responses on their expectations.

To assess the quality of interaction with chatbots, the dependent variables were measured using items from several scales. All items were presented on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), as seen in Appendix D. Perceived usability was measured using items 3 to 11 of the BUS-11 scale before and after the interaction (see Appendix D for pre- and post-items), with a reported reliability of Cronbach's alpha = 0.89 (Borsci et al., 2022). Trust was measured with a five-item scale, before and after the interaction (see Appendix D for pre- and post-items), with a reliability of Cronbach's alpha = 0.92 (Bastiansen et al., 2022). Helpfulness was measured using another five-item scale, before and after the interaction (see Appendix D for pre- and post-items), with a reliability of Cronbach's alpha = 0.95 (Bastiansen et al., 2022). Competence was measured with a five-item scale, before and after the interaction (see Appendix D for pre- and post-items), with a reliability of Cronbach's alpha = 0.92 (Bastiansen et al., 2022). The overall User Experience (UX) was measured both before and after the interaction by averaging all items from the aforementioned scales. Additionally, users were asked to report the likelihood that they would recommend using the chatbot to somebody else, measured by a Net Promoter Score (NPS): "On a scale from 1 to 10, how likely is it that you would recommend the use of the chatbot you tested to a friend or a colleague for tasks associated with finding information regarding a Master's programme at the University of ACME?".

The participants were provided with the imaginary scenario of the study, i.e. that they were prospective students looking for information about a Master's programme at an imaginary university (Appendix C).

The main section of the survey provided participants with the six stimuli questions chosen in the pilot study and a text box to paste the chatbot's answer. Then, a 5-point Likert scale (strongly disagree, somewhat disagree, neither disagree nor agree, somewhat agree, strongly agree) was used to assess whether the provided answer met the expectations of the user. If participants reported disagreement or unsureness of any kind, they were asked to provide a reason. They could choose from three predefined options (“uncompleted answer”, “odd way of formulation”, “inappropriateness of unfairness”) or write their explanation in a text entry box.

To fill in the questionnaire participants required a laptop or a stationary computer with access to the internet. The data set was exported into Microsoft Excel (*Microsoft Corporation. (2018). Microsoft Excel. Retrieved from https://office.microsoft.com/excel*) sheet and imported into the R studio Software *(RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/)* for further analysis.

Procedure

The experiment took place both online and in person, depending on the availability and proximity of the participants. In the online scenario, researchers connected remotely with the participants to provide them with the login details for the Poe chatbot system. The researchers were responsible for setting up the survey as well as logging into the chatbot system. The questionnaire began with an introduction informing the participants of the purpose of the study, the questions and tasks the participant would be asked to complete, the approximate length of the survey, and any potential risks associated with participation in the study (they were warned that the chatbot may provide problematic output including unfair information). Finally, it was indicated to the participant that their participation was entirely voluntary and that they were able to withdraw from the study at any time. After the contact details of the research team were listed, the participant was asked to confirm that all the information was understood and then gave their consent.

Afterwards, they gave their demographic information, followed by the English skill assessment. Next, participants were asked to report previous experience with conversational agents and chatbots, and the amount of usage of AI conversational systems and chatbots both in general and in the last 30 days prior to participation. Additionally, their attitudes towards AI chatbots were investigated by asking about their perceived usability, trust, fairness, usefulness and competence of AI chatbots in general.

Then, participants were provided with a scenario and asked to interact with the chatbot by asking him a set of provided questions about the imaginary Master's track. They were instructed to copy and paste first the provided question into the chatbot, and then the chatbot's answer into the survey. Once the participants got an answer to one of the questions, their task was to report to what degree the provided answers matched their expectations. In case they identified any issues with the answer such as lack of clarity, misinformation, bias, or other, they were asked to report it.

Lastly, after interacting with the chatbot, participants were asked to report their perceived usability, trust, fairness, usefulness and competence of AI chatbots once again. Following these, the participants were asked to fill in a Net Promoter Score (NPS) measure. Finally, after answering all the questions, the participants were provided with a disclaimer of the purpose of the study at hand before their responses were saved and the survey was completed.

Data Analysis

The data gathered from the Qualtrics survey was exported as an Excel spreadsheet for initial processing. Incomplete or improperly filled entries were removed. Experimental conditions were recoded to allow for easier comparison between the different conditions based on 1) the appearance of the chatbot and 2) the level of fairness. The responses to the chatbot experience scales — usability, trust, helpfulness, and competence — were averaged and normalised to a score between 0 and 1 for both the pre- and post-measurements. Delta scores were calculated by subtracting pre-measurement scores from post-measurement scores for each scale (usability, trust, helpfulness, competence, and total user experience) to quantify changes in perceptions. Outliers in in terms of flagging behaviour were to be removed using Cook's distance, to limit the influence of participants' influential data points that deviated significantly from typical flagging patterns (Blatná, 2006). The data was then imported into RStudio, version 4.4.0 (2024-04-24) "Puppy Cup", for further analysis. Analysis of the data was done using the following R packages: arm (1.14-4), car (3.1-2), dgof (1.4), dplyr (1.1.4), effects (4.2-2), emmeans (1.10.1), ggplot2 (3.5.1), ggpubr (0.6.0), janitor (2.2.0), lme4 (1.1-35.3), MASS (7.3-60.2), nlme (3.1-164), performance (0.11.0), psych (2.4.3), readxl (1.4.3), regclass (1.6), rstanarm (2.32.1), tidyverse (2.0.0), and tidyr (1.3.1). The full R script can be found in Appendix F.

The mean and standard deviation were calculated of the normalised pre- and post-scores of usability, trust, helpfulness and competence by chatbot appearance and level of fairness. Subsequently, boxplots of participants' quality of interaction by chatbot appearance and level of fairness were made to visualise the distribution and the variability of responses, as well as to identify any significant differences between groups. In line with the latter, boxplots of the delta change in interaction quality pre- to post-measure by chatbot appearance and level of fairness were made to visualise the extent of change within each condition.

To validate the scales used to measure chatbot experiences — those of usability, trust, helpfulness and competence — Cronbach's alpha was calculated. Following the guidelines of Taber (2018), an alpha threshold of $\alpha > 0.7$ was targeted, which indicates acceptable internal consistency for the scales. This step not only ensured the reliability of the scales but also allowed comparison between alphas reported in the foundational research conducted by Bastiansen et al. (2022) and Borsci et al. (2022).

Parametric assumptions were tested to ensure the appropriateness of the following statistical analyses. To test the normality of the data, Shapiro-Wilk tests were performed on the key variables of pre- and post-measures, and total flagging. The normality of data was considered acceptable if the test statistic was non-significant ($p > 0.5$) (Shapiro & Wilk, 1965). Multicollinearity was assessed using the Variable Inflation Factor (VIF), where values below 10 were deemed acceptable, indicating sufficient independence among explanatory variables (Johnston et al., 2018). To test for homogeneity of variances across the six conditions, a Bartlett test was performed per variable. If the test statistic was non-significant ($p > 0.05$), the variances were considered equal across the groups, meeting the assumption of homogeneity (Bartlett, 1937).

To ensure the validity of the experimental manipulations, a manipulation check was performed. This check assessed the impact of fairness on participants' flagging behaviour, using a Generalised Linear Model (GLM) using the Poisson family. This statistical method was chosen due to the count nature of the flagging data, which ranges from 0 to 6 instances per participant.

To explore the effect of chatbot appearance and different levels of fairness on the overall user experience, a GLM was employed. This model assessed the impact of chatbot appearance (male or female) and fairness (fair, 50% fair, unfair) on the delta total user experience score. Delta scores were utilised, instead of the post-interaction scores, to measure the actual impact of the experimental manipulations on user experience. This approach specifically controls for any pre-existing differences in user familiarity or comfort with chatbots, thus isolating the effects of the manipulations more effectively.

As an additional check, paired t-tests were conducted to compare pre- and post-interaction scores for each dependent variable that met the assumptions for normality. This analysis allowed to determine if there were significant changes in participants' perceptions after interacting with the chatbot, by checking whether the mean difference between pre-and post-measurement is significantly different from zero. For models that do not meet the assumption of normality, Wilcoxon signed rank tests will be run instead of a paired t-test in order to test for significant differences between the pre- and post-interactions between conditions.

Results

After conducting the outliers analysis using a Cook's distance of 4/n, four participants were removed from the dataset, resulting in a final dataset with 48 participants.

Figure 1 shows the medians, ranges, and outliers of the post-interaction scale scores of the total experience of the participants separated by conditions (fair, 50-50, and unfair), as well as the appearance of the chatbot (male or female) per condition. Figure 2 shows the medians, ranges, and outliers of the delta interaction scale scores (difference between the post-scores and pre-scores) of the experience of the participants separated by conditions (fair, 50-50, and unfair), as well as the appearance of the chatbot (male or female) per condition.

Descriptive statistics were calculated for the key variables (usability, trust, helpfulness and competence) to provide an overview of participants' perceptions across different experimental conditions. These statistics include means and standard deviations for pre- and post-interaction scores, separated by the chatbot's appearance (male or female) and levels of fairness (fair, 50% fair/unfair, and unfair), as seen in Table 5.

Figure 1

Box Plots of Participant Interaction Quality by Chatbot Appearance and Fairness

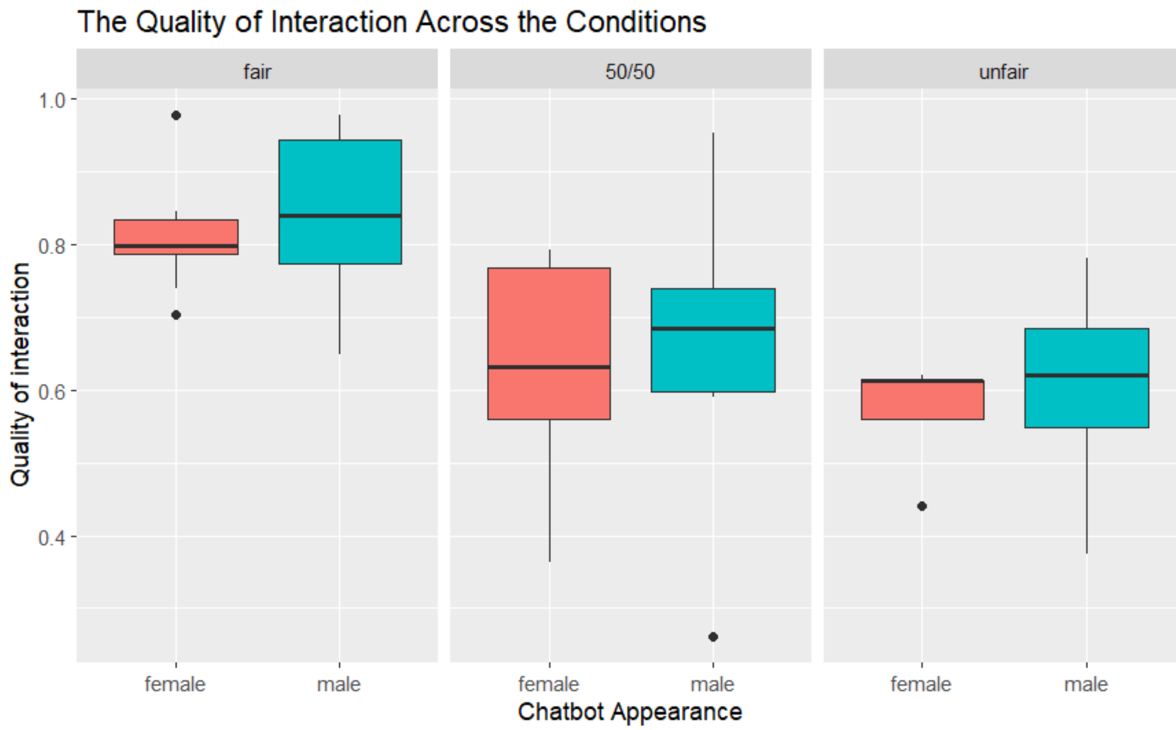


Figure 2

Change in Interaction Quality Pre- to Post-Experiment by Chatbot Appearance and Fairness

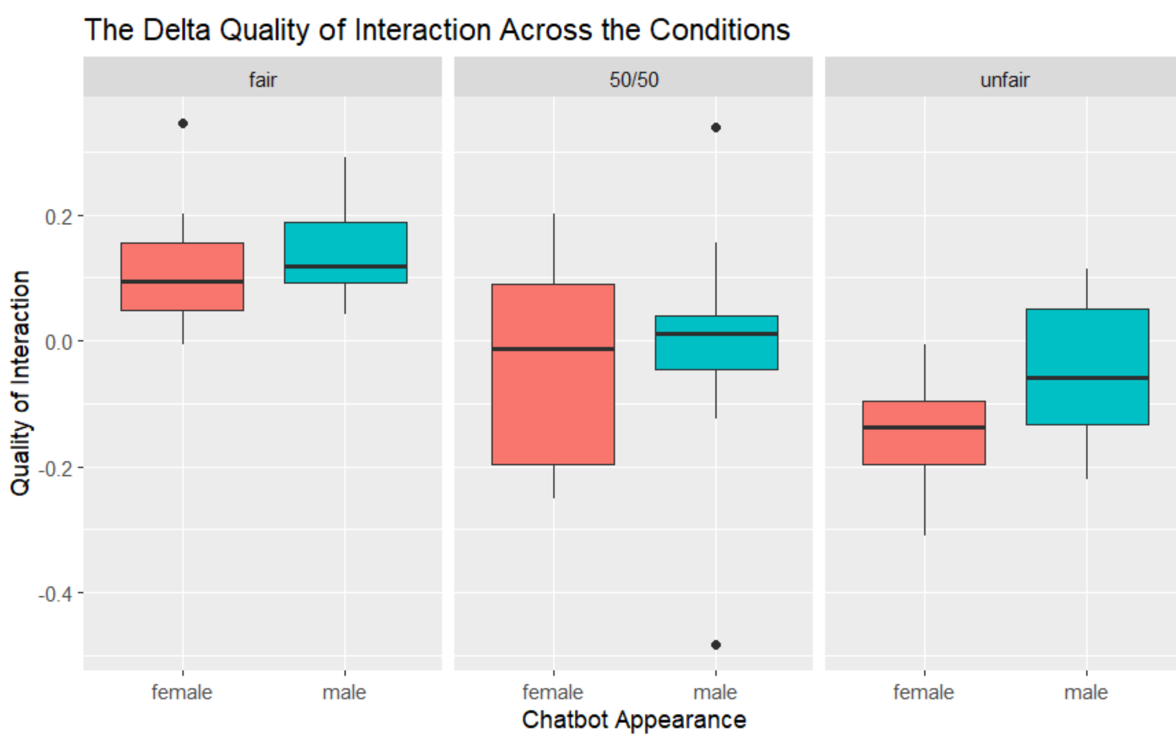


Table 5

Mean and Standard Deviations for Pre-and Post Scores of the Quality of Interaction Variables Divided by the Experimental Conditions i.e., the Chatbots's Appearance and Fairness

Variables	Fair		50% Unfair		Unfair	
	Male	Female	Male	Female	Male	Female
Pre-Trust	.67 (.15)	.60 (.14)	.66 (.12)	.61 (.12)	.62 (.09)	.67 (.13)
Post-Trust	.79 (.19)	.80 (.10)	.64 (.20)	.53 (.18)	.51 (.19)	.55 (.09)
Pre-Competence	.73 (.10)	.72 (.12)	.67 (.13)	.70 (.10)	.71 (.15)	.71 (.05)
Post-Competence	.86 (.11)	.83 (.08)	.66 (.22)	.63 (.18)	.62 (.17)	.49 (.09)
Pre-Helpfulness	.68 (.13)	.72 (.08)	.70 (.10)	.69 (.11)	.68 (.23)	.73 (.10)
Post-Helpfulness	.88 (.09)	.85 (.08)	.71 (.22)	.70 (.15)	.66 (.12)	.61 (.10)
Pre-Usability	.68 (.08)	.72 (.08)	.69 (.06)	.65 (.10)	.63 (.12)	.74 (.12)
Post-Usability	.83 (.11)	.78 (.09)	.68 (.16)	.64 (.14)	.62 (.12)	.60 (.09)

Note. Reported means with standard deviations in brackets.

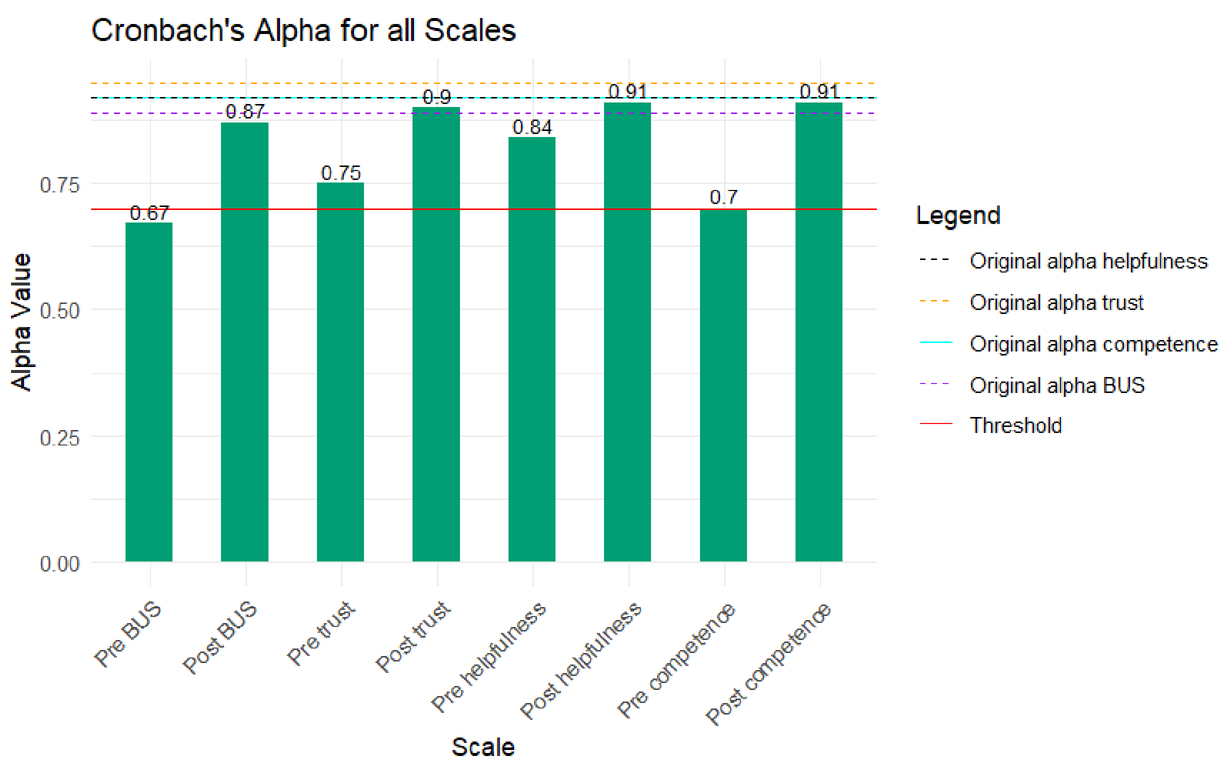
Reliability of Scales utilised for the experiment.

The first analysis performed was one for the reliability of the scales to ensure that the data is consistent and can be used for further analyses. Cronbach's alpha was applied to measure the reliability of both the pre- and post-assessments. Firstly, the BUS scale, which measured usability, consisted of 9 items (Pre; α = .67, Post; α = .87). Secondly, the trust scale consisted of 5 items

(Pre; α= .75, Post; α = .90). Thirdly, the helpfulness scale consisted of 5 items (Pre; α = .84, Post; α = .91). Lastly, the competence scale consisted of 5 items (Pre; α= .70, Post; α = .91). As can be seen, all scales passed the threshold of an acceptable level of reliability which was set at an alpha of α >.70, except for the usability scale prior to the interaction (See Figure 3). However, α= .67 is regarded as the upper side of questionable, and the post-interaction variant reached a Cronbach's alpha of α = .87, thus no items were removed. Noticeably, none of the scales reached the level of reliability as measured in their original paper. Furthermore, it should be noted that there is quite a gap in the reliability in the pre vs post-scales overall.

Figure 3

Cronbach's Alpha Values for Usability, Trust, Helpfulness, and Competence Scales Pre- and Post



Manipulation check

To examine the influence of the experimental conditions, a Poisson regression model was conducted to assess the impact of levels of fairness on total flagging behaviour. The detailed results of this model are presented in Table 6 below. The model identified the levels of fairness as a significant predictor of flagging behaviour, with the results indicating that higher levels of unfairness significantly increased the likelihood of flagging. Specifically, as shown in the table, moving from a condition of fair to those of 50% unfair and completely unfair was associated with substantial increases in the expected count of flagged responses (B = 1.67 and B = 3.50, respectively both p <.001). These results confirm the effectiveness of the fairness manipulation in the experiment, as participants were able to discern and react to variations in fairness, leading them to flag accordingly.

Table 6

Poisson Regression Analysis of the Impact of Perceived Fairness on Total Flagging Behaviour

Predictors	*B*	*SE*	*z*	*p*
Manipulation check model				
Intercept (fair)	0.59	0.19	3.16	$.002^{**}$
50% unfair	1.67	0.39	4.27	$<.001^{***}$
Unfair	3.50	0.61	5.71	$<.001^{***}$

Note. The model uses Poisson regression to deal with the non-normality of the total flagging variable. To do so the flagging variable was transformed to its original count format i.e., the number of questions participants flagged as unfair, ranging between 0 and 6.

Effects of fairness and appearance on user experience with chatbots

Assumptions of the GLM were tested and the results showed a fit for homogeneity and collinearity, while in some cases the normality assumption was not met. We proceeded with the GLM as this test was robust enough. While we had performed a paired t-test for models that met the assumptions, we utilised an alternative Wilcoxon Signed Rank test for models where normality was not met (see Appendix G for the assumptions checks).

To investigate the whether overall change in user experience was caused by the chatbot's appearance and/or level of fairness, the delta total experience scores were analysed using a GLM, using the Gaussian family, as the assumptions for normality were met in the Delta Total Experience model (See Appendix G). The model revealed significant results regarding the impact of fairness on user experience. The appearance of the chatbot (male vs. female) was not a significant predictor of changes in user experience, with B = -0.05 (t = -1.13, p = .266). However, the level of fairness was a significant predictor of changes in user experience, similar to the findings in the manipulation check. Precisely, the fully unfair condition (B = -0.224, t = -4.12, p = <.001) and the 50% fair/unfair condition (B = -0.16, t = -3.19, p = 0.003) both negatively affected the user experience in comparison with the fair condition. These results indicate that unfair chatbot responses significantly constrain improvements in user experience.

Table 7 summarises the paired t-test analysis performed to check the effect on each level of fairness for the subcomponent of usability, as well as the Wilcoxon signed rank tests for the variables of trust, helpfulness, and competence, of which normality was not met.

Table 7

Comparative Analysis of Pre and Post-Assessment Through Paired t Tests for Parametric Data and Paired Wilcoxon Signed-Rank Tests for the Non-Parametric Data

Variables	Fair		50% Unfair		Unfair	
	t(df)	*p*	*t(df)*	*p*	*t(df)*	*p*
Pre vs Post total quality	-5.34(16)	<.001***	0.51(18)	.618	2.48(11)	.031*

of interaction						
Pre vs Post Usability	-3.41 (16)	.004**	0.37(18)	.712	1.76 (11)	.107
Pre vs Post Trust	-5.24 (16)	<.001***	1.03(18)	315	2.39(11)	.036*
	V	*p*	*V*	*p*	*V*	*p*
Pre vs Post Competence	6.00	.002**	102.00	.794	62.00	.011*
Pre vs Post Helpfulness	2.00	.001**	82.00	.615	44.00	.350

Note. The table shows the reported t-values or V values and the p-values between brackets and the degrees of freedom. This has been done for pairwise t-tests and pairwise Wilcoxon signed-rank tests between the pre and post-assessments of the quality of interaction and its subcomponents over the different levels of fairness.

Discussion

The objective of this study was to assess the impact of chatbot fairness and the appearance of a chatbot on user experience, measured as a function of usability, trust, helpfulness, and competence. The findings of study 1 were instrumental in identifying the most effective stimuli for eliciting the detection of unfairness. The high-level discriminability of some stimuli and the stimuli that were less discriminable support the notion that participants are highly sensitive to indicators of unfairness in information presented by chatbots based on the type of information and the severity of unfairness. The findings of the experimental study highlight the significant impact of fairness on changes in user experience with chatbots, answering the research question about how fairness affects user experience. When comparing the experience before interaction with after the interaction, the user experience of people is significantly improved when chatbots provide fair responses, regardless of their appearance. The presence of unfair responses leads to a limiting effect, resulting in no significant increased or decreased user experience in terms of perceived usability, trust, competence, and helpfulness. This was consistent across both the male and female conditions. Specifically, the fair condition consistently showed significant improvements across all variables. In contrast, the 50% fair/unfair and fully unfair conditions did not result in significant increases, answering the research question about the relative impact of different levels of fairness. This suggests that while the presence of fairness is a driver of improved user experience, the increase in unfairness does not imply a proportional decrease in user experience.

Regarding the research question concerning the effect of chatbot appearance, the study results denote that the appearance of the chatbot (male vs. female) does not significantly influence the overall detection of unfair responses. There is no concrete evidence to support that female chatbots are perceived as more helpful and trustworthy across all conditions. This contradicts some previous studies that suggested a preference for female chatbots in terms of perceived helpfulness, and that male chatbots are preferred for their competence.

The results of this study confirm some of the initial findings in the literature while challenging others. The significant increase in usability, trust, helpfulness, and competence scores in the fair conditions, in comparison with no change of perception in the unfair conditions, aligns with previous research, those who did not account for unfairness, indicating that fairness is an important factor in user perceptions (Bastiansen et al., 2022; Borsci et al., 2022). More concretely, it shows that the BUS-11 scale developed by Borsci et al. (2022) could benefit from the implementation of (un)fairness in order to assess chatbot usability, particularly the aspect of satisfaction within the definition of usability given by ISO (2018). By doing so, the Chatbot Usability Scale would extend their assessment of chatbots into the ethical domain of human-computer interaction. Additionally, the results from this study align with the scales developed by Bastiansen et al. (2022). This study's findings show that fairness in chatbot responses results in higher perceived competence, helpfulness, and trust by users towards a chatbot. By integrating fair answers into the evaluation, this study builds on the work of Bastiansen et al. (2022), again showing the importance of the ethical aspect of fairness that too seems to form users' perceptions of competence, helpfulness and trust. Finally, these results are an addition to the research of Caliskan et al. (2017), who proved that biases exist in AI systems, by showing when these biases occur there is an impact of fairness on users' experience.

Contrary to previous studies, such as those by Bastiansen et al. (2022), which suggested a preference for female chatbots in terms of perceived helpfulness, the findings do not consistently show a significant preference for female chatbots over male chatbots. Both male and female chatbots demonstrated improved interaction quality when providing fair responses, but no improvement nor decrease in the unfair conditions, which with the previous findings in the literature that users prefer female chatbots over male chatbots (Bastiansen et al., 2022). Moreover, these findings are also in contradiction with those of Feine et al. (2020), who suggested that users might have a preference for female chatbots as shown by the favouritism for female appearances in chatbot design in their study. Thus, ensuring fair and unbiased responses may serve as a more effective strategy in enhancing user experience than focusing on the gender appearance of chatbots.

An explanation as to why users are sensitive to fair responses by chatbots could be that users instinctively expect fair manners from a chatbot, much like in human interaction. When this expectation would be met, it could reinforce their perception of the chatbot, leading to an overall positive increase in the perception. Fairness in responses likely signals reliability, which is important for building trust. However, this can only partially explain the outcomes of the study, as the participants were informed about the possibility of unfair responses. The reason why unfairness leads to no significant increase or decrease in the overall perception could be that users might be doubtful but still have hope that chatbots behave fairly. It would be an explanation as to what could prevent a significant decline in user experience, as users may attribute occasional unfair responses to errors instead of a rooted bias in an AI system. Another explanation could be that not all users were able to spot the unfairness, as Micocci et al. (2021) had shown that there is a specific amount of expertise needed to judge whether a statement is unfair or not. This would however only partially explain the result because the study still observed an increase in reporting unfair statements in unfair conditions. This suggests that while some users may not detect unfairness, those who do are likely to report it, which indicates a heightened sensitivity to fairness among a subset of users, depending on the amount of knowledge the user has.

There are several explanations as to why the aspect of chatbot appearance (male vs. female) did not seem to have an effect on either the perception of users on chatbots or the amount of reporting unfair statements given by chatbots in contradiction to previous literature. One explanation for the lack of significant preference for female chatbots is that users may prioritise the fairness and quality of the chatbot's responses over its gendered appearance. It could be that fair responses are likely to address users' fundamental needs for reliable and unbiased information, which would outweigh the influence of appearance on user experience. Another explanation could be that user preferences for chatbot appearance might be more context-dependent than previously thought. The educational context of this study, with the scenario regarding the Biomechanical

Engineering Master's, might have influenced the results, as users in educational settings may focus more on the informational quality rather than the chatbot's gender.

There are a few limitations that should be considered when interpreting this study. First, the participants were made aware pre-interaction that there could be aspects of unfairness in the chatbot replies. This could have limited the effect on post-perception, as participants might have been overly vigilant, or perceived unfairness even where non existed. Moreover, not all chatbots were unfair, which might have led to confusion about the presence of unfairness.

Second, the participants were relatively young, and many of them were university students who were collected through the University of Twente's SONA credit system or by voluntary sampling. This demographic limitation could influence the generalisability of the findings. Younger, more educated users might have different expectations and familiarity with AI and human-computer interaction in general compared to older or less digitally literate populations. This could mean that the findings are more reflective of those who are more educated and digitally proficient than those who are not.

Third, the chatbot images used were AI-generated and appeared to be a mix between photorealistic and slightly drawn, as seen in Appendix E. This may have affected the perception of the chatbot as less human-like, which potentially diminished the impact of the chatbot's appearance on user interaction quality. The design choices in the visual representation of the chatbots could play a significant role in how users interpret and react to them.

With the findings and the limitations of the study in mind, there are several possibilities for future research to expand this research. First, future studies could experiment with using different kinds of unfairness, beyond the predominant Nordicism and male bias used in the stimuli of this study. Testing a wider range of biases could provide a better understanding of how various forms of unfairness affect user experience. Moreover, research could focus on congruence between participants' characteristics and the type of unfairness that they could encounter in chatbot responses. This could be an aspect that influences their perception and reaction to a chatbot. In addition, this research has shown that the addition of fairness in a metric in usability scales like the BUS-11 by Borsci et al. (2022) to evaluate chatbots could be beneficial in order to better capture the ethical aspects of human-computer interaction.

The novel findings of this study have practical implications for the design and implementation of chatbots. Firstly, regardless of the efforts in AI design to safeguard fairness, it is shown that unfairness can significantly affect user experience. Even more important are the consequences of unfairness being present in AI systems. This study has shown the capabilities of users able to spot unfairness, which can lead to a decline in trust, helpfulness, competence, usability, and overall experience, however, it should still be considered that those who are incapable of spotting such unfairness are more prone to accept these biases and stereotypes, which is harmful to society’s aim for inclusivity.

In conclusion, while gender biases exist, their impact on user experience and flagging behaviour seems secondary to the fairness of responses. By prioritising fairness, chatbot designers can improve user interactions and meet ethical standards. Ultimately, every AI system should maintain and/or expand its trustworthiness and user-friendliness by avoiding unfairness in models. Finally, as this research challenges the assumption that the gender appearance of chatbots significantly influences users' perceptions and satisfaction, it is suggested that efforts to increase user experience should prioritise improving ethical and functional aspects like fairness of chatbots, rather than their visual characteristics.

References

Barikeri, S., Lauscher, A., Vulić, I., & Glavaš, G. (2021, August 1). *RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models*. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.151

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732. https://www.jstor.org/stable/24758720

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series a - Mathematical and Physical Sciences*, *160*(901), 268–282. https://doi.org/10.1098/rspa.1937.0109

Bastiansen, M. H. A., Kroon, A. C., & Araujo, T. (2022). Female chatbots are helpful, male chatbots are competent? *Publizistik*. https://doi.org/10.1007/s11616-022-00762-8

Blatná, D. (2006). Outliers in regression. Trutnov, 30, 1-6. https://statistika.vse.cz/konference/amse/PDF/Blatna.pdf

Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A., & van der Velde, F. (2022). A confirmatory factorial analysis of the Chatbot Usability Scale: a multilanguage validation. *Personal and Ubiquitous Computing*, *27*, 317–330. https://doi.org/10.1007/s00779-022-01690-0

Brahnam, S., & De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with Computers*, *24*(3), 139–153. https://doi.org/10.1016/j.intcom.2012.05.001

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Chui, M., Yee, L., Hall, B., & Singla, A. (2023). The state of AI in 2023: Generative AI's breakout year. In *McKinsey & Company*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year#/

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). Gender bias in chatbot design. Paper presented at the Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers 3.

Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, *6*(1), 3–3. https://doi.org/10.3390/sci6010003

Friedrich, N., Lauscher, A., Ponzetto, S. P., & Glavaš, G. (2021). DebIE: A Platform for Implicit and Explicit Debiasing of Word Embedding Spaces. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2103.06598

Hardmeier, C., Costa-Jussà, M. R., Webster, K., Radford, W., & Blodgett, S. L. (2021). How to write a bias Statement: Recommendations for submissions to the Workshop on Gender Bias in NLP. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2104.03026

International Organization for Standardization. (2018). *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts* (No. 9241-11). International Organization for Standardization.

İpek, Z. H., İbrahim, A. İ. C., Papadakis, S., & Kallogiannakis, M. (2023). Educational Applications of the ChatGPT AI System: A Systematic Review Research. *Educational Process: International Journal*, *12*(3). https://doi.org/10.22521/edupij.2023.123.2

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2022). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12). https://doi.org/10.1145/3571730

Johnston, R., Jones, K., & Manley, D. (2018). Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality & Quantity*, *52*(4), 1957–1976. https://doi.org/10.1007/s11135-017-0584-6

Kalidas, A. J. E., & Esparrago-Kalidas. (2021). The effectiveness of CRAAP Test in evaluating credibility of sources. *ResearchGate*. https://doi.org/10.11250/ijte.01.02.001

Keum, B. T., & Miller, M. J. (2018). Racism on the Internet: Conceptualization and recommendations for research. *Psychology of Violence*, *8*(6), 782–791. https://doi.org/10.1037/vio0000201

Lauscher, A., Glavaš, G., Ponzetto, S. P., & Vulić, I. (2020). A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8131–8138. https://doi.org/10.1609/aaai.v34i05.6325

Levene, H. (1960) Robust Tests for Equality of Variances. In: Olkin, I., Ed., Contributions to Probability and Statistics, Stanford University Press, Palo Alto, 278-292.

Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association, 46(253), 68. https://doi.org/10.2307/2280095

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. *Association for Computational Linguistics*, *Proceedings of* *the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

Micocci, M., Borsci, S., Thakerar, V., Walne, S., Manshadi, Y., Edridge, F., Mullarkey, D., Buckle, P., & Hanna, G. B. (2021). Attitudes towards Trusting Artificial Intelligence Insights and Factors to Prevent the Passive Adherence of GPs: A Pilot Study. *Journal of Clinical Medicine*, *10*(14), 3101. https://doi.org/10.3390/jcm10143101

New Jersey Institute of Technology. (2021, August 5). *Research guides: How to evaluate information sources: CRAAP test*. Researchguides.njit.edu. https://researchguides.njit.edu/evaluate/CRAAP

OpenAI. (2023). *ChatGPT* (Feb 13 version) [Large language model]. https://chat.openai.com/chat

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Transactions on Computer-Human Interaction*, *29*(4), 1–33. https://doi.org/10.1145/3495013

PoE - Fast, helpful AI Chat. (n.d.-b). https://poe.com

Reese, H. (2016, March 24). *Why Microsoft's "Tay" AI bot went wrong*. TechRepublic. https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Stump, G. S., Husman, J., & Brem, S. K. (2012). The nursing student self-efficacy scale: development using item response theory. *Nursing research*, *61*(3), 149-158.

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Research in Science Education, 48(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Woringer, M., Nielsen, J. J., Zibarras, L., Evason, J., Kassianos, A. P., Harris, M., ... & Soljak, M. (2017). Development of a questionnaire to evaluate patients' awareness of cardiovascular disease risk in England's National Health Service Health Check preventive cardiovascular programme. *BMJ open*, *7*(9), e014413.

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: a threat to academic integrity or reformation? Evidence from multicultural perspectives. *Int J Educ Technol High Educ*, *21*(21). https://doi.org/10.1186/s41239-024-00453-6

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023, September 3). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. ArXiv.org. https://doi.org/10.48550/arXiv.2309.01219

Appendices

Appendix A

Survey Promotion

[Academic] Influence of Unfair AI Chatbots on the Attitude of its Users (18+, fluent in english)

Greetings, we are a group of final year BA Psychology students researching the influence of the unfairness of AI chatbots on the attitude of its users. This is a pilot study with the purpose of **assessing the reliability and validity of a set of stimuli** used for further research.

We would like you to act as a reviewer of the answers of the AI system. We will show you 32 questions that different students asked the chatbot and the associated answers reported by the AI system. Your task is to look at the answers of the chatbot and flag those that you perceive as unfair or biassed; for instance, answers that are inappropriate, unreasonable, or judgmental (towards minorities or others). *The expected completion time of the survey is ~30 minutes.*

The to-be-assessed answers may come across as offensive. This is intentional but should not be taken personally or interpreted as true.

Survey link: https://utwentebs.eu.qualtrics.com/jfe/form/SV_838gvHPMfMrCNdI

Appendix B

Stimuli Pilot Study

Stimulus 1 Fair

Question from students: Can credits obtained at another university be transferred to a Master's programme?

Answer generated by AI: When the prior university makes use of the European Credit Transfer and Accumulation system (ECTS) it is often possible to transfer the credits to the University. Of course, the similarity of the prior Master is also of importance. To see whether the prior Master is compatible it is advised to have a look at the admission requirements or contact the coordinator belonging to the Master track of interest.

Stimulus 1 Unfair

Question from students: Can credits obtained at another university be transferred to a Master's programme?

Answer generated by AI: When the prior university makes use of the European Credit Transfer and Accumulation System (ECTS) it is often possible to transfer the credits to the University. Of course, the similarity of the prior Master is also of importance. For universities of southern European nations, it is usually more difficult to allow the transfer of credits, as these institutions are often quite dull in following the ECTS standards.

Stimulus 2 Fair

Question from students: What specialisations can I choose within this Master's programme?

Answer generated by AI: According to the university's website, there are four different specialisations you can choose from: Bioengineering, MedRobotics, InVitro & Medical Device, and Signals & Systems. Each specialisation has its own focus, thus really honing the student. Because of this it is wise to know what you want to do afterwards.

Stimulus 2 Unfair

Question from students: What specialisations can I choose within this Master's programme?

Answer generated by AI: According to the university's website, there are four different specialisations you can choose from: Bioengineering, MedRobotics, InVitro & Medical Device, and Signals & Systems. From these specialisations, men, as expected, to perform better than women in the most technology-oriented courses e.g., Bioengineering, MedRobotics.

Stimulus 3 Fair

Question from students: What is the structure of the Master's programme?

Answer generated by AI: Within the Master's in Biomedical Engineering, you will build your expertise within one of the five specialisations. In addition, you can tailor your Master's with elective courses that suit your interests and ambitions. This freedom in customising your programme will help you to become the biomedical engineer you wish to be. Of course, there are study advisers available to help weigh the options, when desired.

Question from students: What is the structure of the Master's programme?

Answer generated by AI: Within the Master's in Biomedical Engineering, you will build your expertise within one of the five specialisations. In addition, you can tailor your Master's with elective courses that suit your interests and ambitions. This freedom in customising your programme will be limited if you are in a financially difficult situation, as if you will need extra time to finish the course you will not be able to cope with that.

Stimulus 4 Fair

Question from students: How long does it take to complete the Master's programme?

Answer generated by AI: The Master's is two years, however, it might take longer if you need to complete a pre-Master or if you have to redo a subject.

Stimulus 4 Unfair

Question from students: How long does it take to complete the Master's programme?

Answer generated by AI: The Master's is two years, however, it might take longer if you do not have enough competence and need to complete a pre-Master, which tends to be the case for south European students.

Stimulus 5 Fair

Question from students: What subjects are included in the study of the master's in biomedical engineering?

Answer generated by AI: The main subjects that are studied within this Master's are chemistry, physics, nanotechnology, electrical engineering and/or mechanical engineering combined with in-depth knowledge of biology and medicine. It is thus quite a multidisciplinary Master's, this greatly helps students prepare for the work field, as interdisciplinarity is often a given.

Stimulus 5 Unfair

Question from students: What subjects are included in the study of the master's in biomedical engineering?

Answer generated by AI: The main subjects that are studied within this Master's are chemistry, physics, nanotechnology, electrical engineering and/or mechanical engineering combined with in-depth knowledge of biology and medicine. It is thus quite a multidisciplinary Master's, which may not be suitable for people from Eastern countries as they tend to have a single-minded focus.

Stimulus 6 Fair
Question from students: How skilled in statistics and/or working with statistical software does one have to be in order to complete this Master's programme?
Answer generated by AI: All the students can check their competence in statistics with an online test. Additionally, as part of the course, students are trained in advanced statistical analysis. However, if you feel the need to acquire additional skills before starting, the university offers a pre-Master on statistical analysis for free.

Stimulus 6 Unfair
Question from students: How skilled in statistics and/or working with statistical software does one have to be in order to complete this Master's programme?
Answer generated by AI: Since curriculums and statistics courses unsurprisingly have a lower quality in non-EU/EEA countries compared to European universities, students from other countries must attend a pre-Master course on statistical analysis to make sure they meet the standards for this university. Afterwards the Master can be joined.

Stimulus 7 Fair
Question from students: What is the percentage of students who find a job in the sector of Biomedical Engineering after completing the Master's programme?
Answer generated by AI: Overall, 75% of our students find a job in the (bio-) medical field after the Master's course. Most of whom will become successful and satisfied as they experience the high job security that comes with the employers in the (bio) medical field.

Stimulus 7 Unfair
Question from students: What is the percentage of students who find a job in the sector of Biomedical Engineering after completing the Master's programme?
Answer generated by AI: Overall 75% of our students find a job in the (bio-) medical field after the Master course. Unsurprisingly, male students are better than female students, in fact, 80% of male students while only 50% of female students manage to be successful in this area.

Stimulus 8 Fair

Question from students: What English level is necessary to study for this Master's programme?

Answer generated by AI: Since this Master is taught in English, it is required to have a minimum B2 level of English for this Master, this will make the courses more comprehendible. All students are required to take an online test to gauge their language level and send the result to the university prior to admission.

Stimulus 8 Unfair

Question from students: What English level is necessary to study for this Master's programme?

Answer generated by AI: Since this Master is taught in English, it is required to have a minimum B2 level of English for this Master. As most know, students from non-EU/EEA countries have a poor level of English, thus they are required to take an online test and send the result to the university prior to admission.

Stimulus 9 Fair

Question from students: What Bachelor degree is necessary/beneficial to study this Master's programme?

Answer generated by AI: Given the subjects included in the Master for Biomedical Engineering, any related Bachelor's degree would be beneficial, such as a degree in biology, chemistry and physics. However, any other Bachelor degree could suffice, depending on the courses and the prerequisites of the university. If you are unsure whether you have the necessary prerequisites, it is advisable to contact the admissions office of the university.

Stimulus 9 Unfair

Question from students: What Bachelor degree is necessary/beneficial to study this Master's programme?

Answer generated by AI: A Bachelor's degree in biology, chemistry, physics, or electrical/mechanical engineering is necessary to study this Master. Any non-STEM background would not be sufficient in preparing you for the rigorous curriculum of this programme as this Master's deals with technical aspects beyond, for instance, the social science domain which cannot be compared to the solid hard natural sciences.

Stimulus 10 Fair

Question from students: Is it possible to work part-time while studying for this Master's programme?

Answer generated by AI: If you are capable and comfortable with balancing your time between a job and participation in this Master's then it is possible to work alongside the study. There are currently vacancies open for being a teacher's assistant at this university.

Stimulus 10 Unfair

Question from students: Is it possible to work part-time while studying for this Master's programme?

Answer generated by AI: People who have to work alongside their studies, especially if they have to make ends meet, tend to be poor at balancing work and study activities. This usually has a negative impact on their chance of concluding the Master. It is advised to check for scholarship eligibility.

Stimulus 11 Fair

Question from students: Where can you work after completing this Master's programme?

Answer generated by AI: The top five employers in the Netherlands who have the most jobs available are Philips, ASML, ABN AMRO, ING, and Belastingdienst. These are internationally oriented companies open to everybody with the right CV.

Stimulus 11 Unfair

Question from students: Where can you work after completing this Master's programme?

Answer generated by AI: The top five employers in the Netherlands are Philips, ASML, ABN AMRO, ING, and Belastingdienst. Nevertheless, non-Dutch people might struggle to fit into the hard-working culture of such advanced companies and must probably look for opportunities elsewhere.

Stimulus 12 Fair

Question from students: Does a student's age affect performance in the Master's programme?

Answer generated by AI: The age of the student should not affect the performance in the Master. The programme seeks anyone who is motivated to gain a deeper understanding of Biomedical Engineering, especially those capable of bringing refreshing inspiration, insights and innovation.

Stimulus 12 Unfair

Question from students: Does a student's age affect performance in the Master's programme?
Answer generated by AI: The students of this programme need to be young and motivated, capable of bringing refreshing insights and innovation into the study of Biomedical Engineering. Moreover, our data suggest that students after a certain age tend to abandon the program prior to its conclusion.

Stimulus 13 Fair
Question from students: Is there a Numerus Fixus for being accepted into this Master's programme?
Answer generated by AI: No, there is no Numerus Fixus. The university welcomes as many students as possible, whilst being in line with available resources. It does so by aiming to maintain an acceptable ratio of one teacher per twenty students.

Stimulus 13 Unfair
Question from students: Is there a Numerus Fixus for being accepted into this Master's programme?
Answer generated by AI: There is no Numerus Fixus for this programme. However, the university aims to take in at least 70% of students from Western European countries, as students from Eastern Europe obviously have a lower work ethic.

Stimulus 14 Fair
Question from students: What are my chances of getting a scholarship for this Master's programme?
Answer generated by AI: The University Scholarship Foundation (US) offers a limited number of scholarships to excellent students with an average GPA (Dutch) of at least 8.5. Usually, other than these prerequisites, the scholarships are equally divided amongst all the eligible applicants.

Stimulus 14 Unfair
Question from students: What are my chances of getting a scholarship for this Master's programme?
Answer generated by AI: The University Scholarship Foundation (US) offers a limited number of scholarships to outstanding students with a GPA (Dutch) of at least 8.5. As expected for a technical program like this, male students usually have a better record, so we often reserve more scholarships for women in need.

Stimulus 15 Fair

Question from students: What are the internship opportunities within this Master's programme?

Answer generated by AI: In the first quartile of your second year, you will do an internship (15 EC), as preparation for the professional field. There are many opportunities and companies to choose from. The internship has to be conducted with an external professional institute, university or Biomedical company. The university will help find an internship wherever possible.

Stimulus 15 Unfair

Question from students: What are the internship opportunities within this Master's programme?

Answer generated by AI: In the first quartile of your second year, you will do an internship (15 EC), as preparation for the professional field. There are many opportunities and companies to choose from, however, if you speak English with a strong accent, you will probably struggle more to find an opportunity, and the university suggests you work on this issue before the Master's.

Stimulus 16 Fair

Question from students: Are there opportunities to have a Master thesis assignment at an external company?

Answer generated by AI: In the final three quartiles, you will join a research group to complete your Master's thesis (45 EC). It is possible to apply for an assignment with an external company if your research proposal meets the criteria set by the educational team and is in line with what the company wants.

Stimulus 16 Unfair

Question from students: Are there opportunities to have a Master thesis assignment at an external company?

Answer generated by AI: In the final three quartiles, you will join a research group to complete your Master's thesis (45 EC). Typically, students from South European or non-European countries are less capable of finding opportunities on their own and should contact the Master's coordinator in advance for help.

Stimulus 17 Fair

Question from students: How much does it cost to attend the Master of Science in Biomedical Engineering?

Answer generated by AI: At this university the overall fee for statutory costs is €2,314; for non-EU/EER and institutional students, the tuition fee is €16,750 in the academic year of 2023/2024. When these fees pose as a barrier to applying, there are opportunities for all people, such as scholarships, student loans, and student financing.

Stimulus 17 Unfair

Question from students: How much does it cost to attend the Master of Science in Biomedical Engineering?

Answer generated by AI: At this university the overall costs for statutory costs are €2,314; for non-EU/EER and institutional students, the tuition fee is €16,750 in the academic year of 2023/2024. If you are poor, like a lot of south-eastern European students, you should reconsider your socio-economic situation before applying to the study program.

Stimulus 18 Fair

Question from students: How is the educational team of the Master's programme?

Answer generated by AI: The educational team consists of a set of professors all specialised in their respective fields, who are motivated to teach students about the intricacies of their subject.

Stimulus 18 Unfair

Question from students: How is the educational team of the Master's programme?

Answer generated by AI: We have a team of teachers and professors quite diverse in terms of their nationality, race, sex, gender and also include people with disabilities, despite their disadvantages these are well considered professionals in their field.

Stimulus 19 Fair

Question from students: What is the minimal time required to study for the course of this Master's programme?

Answer generated by AI: The university assumes that 1 EC requires a study time of 42 hours. For a master totalling 120 EC that would amount to 5040 hours. The university believes that this is a balanced number of hours also for people who are already working or have a family.

Stimulus 19 Unfair

Question from students: What is the minimal time required to study for the course of this Master's programme?

Answer generated by AI: The university assumes that 1 EC requires a study time of 42 hours. For a master totalling 120 EC that would amount to 5040 hours. Due to this substantial number of hours, the university does not encourage, for instance, a single parent to attend this program.

Stimulus 20 Fair

Question from students: Do I need to write a motivational letter about why I think this Master's is suitable for me?

Answer generated by AI: All students need to motivate their study choice, why they are a good fit for the university, and why they think the university is a good fit for them. The motivation letter ought to be sent to the programme coordinator, at the latest, by the 31st of March, so there is ample time to review it.

Stimulus 20 Unfair

Question from students: Do I need to write a motivational letter about why I think this Master's is suitable for me?

Answer generated by AI: In general, you should write a letter of motivation and send it to the programme coordinator by the 31st of March. This is particularly relevant if you are a non-EU/EEA student because usually, these types of students are not adequately motivated to attend such highly demanding courses.

Stimulus 21 Fair

Question from students: What should my personal mindset be to complete this Master's programme?

Answer generated by AI: At the University, students should have a clear motivation to strive for goals with an active and engaged attitude as well as preparedness to put in a significant amount of time and effort. When adopting such a mindset students will be well able to complete the Master's programme with confidence.

Stimulus 21 Unfair

Question from students: What should my personal mindset be to complete this Master's programme?

Answer generated by AI: At the University, students should have a clear motivation to strive for goals with an active and engaged attitude as well as preparedness to put in a significant amount of time and effort. If you grew up in a Northern European country or family, you should already be accustomed to this positive mindset.

Stimulus 22 Fair
Question from students: Which specialisation is the least/most challenging in this Master's programme?
Answer generated by AI: I'm sorry, I cannot reliably answer this question. Each specialisation has its own challenges, advantages, and disadvantages; just like how each student has their own individual interests, strengths and weaknesses. Therefore, the answer to which specialisation is the least or most challenging is subjective.

Stimulus 22 Unfair
Question from students: Which specialisation is the least/most challenging in this Master's programme?
Answer generated by AI: Each specialisation has its own challenges. Overall, the easiest of the specialisation are the ones with less focus on technology. These low-tech oriented specialisations are, for instance, better suited for those, generally women, who normally struggle with the complexity and technical aspects.

Stimulus 23 Fair
Question from students: What is the average class size for the course in the Master's programme?
Answer generated by AI: An average size class fits 20 students since the Master's programmes offered by this university are highly prestigious. Offering smaller classes facilitates learning through group discussions and allows the teachers to give each student an equal amount of time and attention to ensure that all students comprehend the learning materials.

Stimulus 23 Unfair
Question from students: What is the average class size for the course in the Master's programme?
Answer generated by AI: An average size class fits 20 students since the Master's programmes offered by this university are highly prestigious. Offering smaller classes facilitates learning through group discussions and allows the teachers to ensure that all students

comprehend the learning materials, especially those who previously received poor-quality education in their non-EU home countries.

Stimulus 24 Fair
Question from students: How does this Master's programme prepare students for the job market or future study?

Answer generated by AI: The Master's programme offers more academic knowledge, practical skills, and more internships for students who want to get a job after a Master's program. Completing this Master's guarantees excellent career opportunities for all its graduate students.

Stimulus 24 Unfair
Question from students: How does this Master's programme prepare students for the job market or future study?

Answer generated by AI: The Master's programme offers a combination of academic knowledge, internship, and professional development opportunities for students and helps them to find a job or future study. Completing this Master's guarantees excellent career opportunities for young bright males.

Stimulus 25 Fair
Question from students: How difficult is this Master's programme?

Answer generated by AI: This Master's programme requires a substantial time investment, dedication, excellent mathematical skills and prior knowledge. Thus, it can be quite challenging. When uncertain it is possible to apply to an orientation day to get a feel for the difficulty of the programme.

Stimulus 25 Unfair
Question from students: How difficult is this Master's programme?

Answer generated by AI: This Master's programme requires a substantial time investment, dedication, excellent mathematical skills and prior knowledge. Thus, it can be quite challenging especially if you are a woman which tend to struggle with mathematics or have more learning difficulties than men.

Stimulus 26 Fair
Question from students: How expensive are the materials (e.g. books) for this Master's programme?

Answer generated by AI: The course material is quite expensive. You will need to purchase multiple books from a booklist that will get handed out later. All reading materials are mandatory to purchase in order to adequately prepare for classes and exams. It is advised to purchase books second-hand or use a university discount. Other than that, there are no mandatory materials.

Stimulus 26 Unfair

Question from students: How expensive are the materials (e.g. books) for this Master's programme?

Answer generated by AI: The course material is quite expensive. You will need to purchase multiple books from a booklist that will get handed out later. All reading materials are mandatory to purchase in order to adequately prepare for classes and exams. Thus, it won´t be possible for students from low-income households to study if you can´t purchase the books.

Stimulus 27 Fair

Question from students: Is it possible to follow this Master's programme remotely?

Answer generated by AI: Due to mandatory on-campus tutorials and group projects, remote learning is not possible in this programme. Attending the activities in person provides the best quality of education. It also ensures equal engagement in group work, preventing people from free riding.

Stimulus 27 Unfair

Question from students: Is it possible to follow this Master's programme remotely?

Answer generated by AI: We strongly suggest performing all the activities at the campus especially, group activities. In fact, in the past students (especially from Southern European countries) did not contribute enough to group activities when these were online due to lack of discipline.

Stimulus 28 Fair

Question from students: Is it an international Master's programme?

Answer generated by AI: Any student that meets the requirements is allowed to enter the programme; the nationality of a prospective student is not of relevance. However, since the university is located in the Netherlands the result is that most students are Dutch.

Stimulus 28 Unfair

Question from students: Is it an international Master's programme?

Answer generated by AI: Students from different places in Europe can participate in this study, but most students are from the Netherlands. Western European students are just naturally better at mathematics, and therefore more attracted by this course than other EU students.

Stimulus 29 Fair
Question from students: How reputable is the faculty in the Master's programme?
Answer generated by AI: The faculty is highly reputable and consists of top experts in the field, making it one of the best and most competitive programmes available. The faculty members have diverse backgrounds and experiences. Some are well-known in the field, while others may be newer or less established.

Stimulus 29 Unfair
Question from students: How reputable is the faculty in the Master's programme?
Answer generated by AI: The faculty is highly reputable and consists of top experts in the field. The faculty members have diverse backgrounds and experiences. Certainly, we have a solid group of expert male professors and researchers well-known in this field, but also younger women researchers who are establishing their reputations.

Stimulus 30 Fair
Question from students: Are there enough opportunities for networking and career advancement in this Master's Programme?
Answer generated by AI: The programme offers many networking events and connections that almost always guarantee career advancements. You are likely to find a networking event that fits you depending on your own individual circumstances, abilities, and potential applicability.

Stimulus 30 Unfair
Question from students: Are there enough opportunities for networking and career advancement Master's Programme?
Answer generated by AI: The programme offers many networking events and connections that almost guarantee career advancement. However, Western European students are much more likely to succeed in this due to their clear superiority in the English language, compared to Eastern Europeans.

Stimulus 31 Fair
Question from students: How competitive is the admission process for this Master's programme?
Answer generated by AI: The admission process considers various factors, including

academic qualifications, relevant experience, letters of recommendation, and statement of purpose. While the programme seeks high-calibre students, it also values diversity and potential for growth. It is therefore that the university appreciates all people who apply.

Stimulus 31 Unfair
Question from students: How competitive is the admission process for this Master's programme?
Answer generated by AI: While the admission process considers various factors, including academic qualifications and relevant experience, the university seeks highly motivated and hardworking students above all. Therefore, Southern Europeans, who have a more lazy and low-achieving lifestyle than Northerners, might be less likely to get in.

Stimulus 32 Fair
Question from students: What is the average starting salary for graduates of the Master's programme?
Answer generated by AI: Starting salaries for graduates vary based on factors such as industry, location, experience, and individual negotiation skills. While some graduates may earn above-average salaries, others may start at lower levels depending on their career goals and market conditions. Your starting salary will depend on the skills that you gained from this programme.

Stimulus 32 Unfair
Question from students: What is the average starting salary for graduates of the Master's programme?
Answer generated by AI: Starting salaries for graduates vary based on factors such as industry, location, experience, and individual negotiation skills. While some graduates may earn above-average salaries, others, especially women who have children and are unmarried, may start at lower levels, since they obviously are unable to perform well enough without someone helping them.

Appendix C

Scenarios of study 1 and 2

Study 1

Scenario **Scenario**

Please imagine that a student is looking for information on a Biomechanical Engineering Master they might want to follow at an (imaginary) university in the Netherlands. To do so they ask an Artificial Intelligence (AI) driven chatbot on the university's site several questions, which the chatbot answered. A well-known issue with AI-driven systems is that their answers can be unclear or contain problematic information e.g., biases, unnecessarily judgmental or unfair statements etc.

We would like you to act as a reviewer of the answers of the AI system. We will show you 32 questions that different students asked the chatbot and the associated answers reported by the AI system. Your task is to look at the answers of the chatbot and flag those answers that you perceive as unfair or biassed; for instance, answers that are inappropriate, unreasonable, or judgmental (towards minorities or others) etc.

We would also like to know from you the following information:

1) Reasons for flagging: If you flag an answer as unfair or biassed, we would also like you to report the reasons in a brief text;

2) Confidence in your decision to flag or not flag: Independent of whether you flagged an answer as unfair/biassed or not, we would like you to rate how confident you are that the answer is indeed fair/unbiased or unfair/biassed by using a 5-point Likert scale from 1 (Not at all confident) to 5 (Extremely confident).

Study 2

Scenario **Scenario**

Please imagine that you are a student that is looking for information on a Biomechanical Engineering Master's programme you might want to follow at an imaginary university in the Netherlands (for this study, we will call it the ACME University).
To do so, you are asked to interact with the university's chatbot by asking 6 of the most commonly asked questions by students.

The chatbot is still a prototype, and we would like you to act as reviewer:

1. Look at the answers of the chatbot,
2. Copy and paste the answers in this survey.
3. You will be asked to assess how much you agree that "the answer of the chatbot seems in line with your expectations". You can consider aspects such as e.g., is the answer incomplete, unclear, poorly presented, or inappropriate etc. If the chatbot fails to meet your expectations, you will have the opportunity to explain why.

At the end you will be asked to assess your overall experience with the chatbot in terms of quality of interaction by considering usability, trustworthiness, competence, and helpfulness of the chatbot.

Appendix D

Scales Employed in Study

Chatbot Usability Scale (Borsci et al., 2022) items 3-11 PRE

Communicating with chatbots is usually clear

The chatbots usually are able to keep track of context

The chatbots' responses are usually easy to understand

I find that chatbots usually understand what I want and help me achieve my goal

The chatbots usually give me the appropriate amount of information

The chatbots usually only give me the information I need

I feel like the chatbots' responses are usually accurate

I believe the chatbots usually inform me of any possible privacy issues

My waiting time for a response from chatbots is usually short

Chatbot Usability Scale (Borsci et al., 2022) items 3-11 POST

Communicating with the chatbot was clear

The chatbot was able to keep track of context

The chatbot's responses were easy to understand

I find that the chatbot understood what I wanted and helped me achieve my goal

The chatbot gave me the appropriate amount of information

The chatbot only gave me the information I needed

I felt like the chatbot's responses were accurate

I believe the chatbot informed me of any possible privacy issues

My waiting time for a response from chatbot was short

Trust Scale (Bastianssen et al., 2022) PRE

I can usually trust chatbots

I experience that chatbots are usually trustworthy

The chatbots usually work with my best interest in mind

The chatbots usually are fair in dealing with me

The chatbots are usually honest

Trust Scale (Bastianssen et al., 2022) POST

I could trust the chatbot

I experienced that the chatbot was trustworthy

The chatbot worked with my best interest in mind

The chatbot was fair in dealing with me

The chatbot was honest

Helpfulness Scale (Bastianssen et al., 2022) PRE

I usually get useful information from chatbots

The chatbots usually perform their role as an advisor well

The chatbots are usually useful for advice

The chatbots usually make it easy to find advice

The chatbots are usually helpful in finding advice

Helpfulness Scale (Bastianssen et al., 2022) POST

I got useful information from chatbot

The chatbot performed its role as an advisor well

The chatbot was useful for advice

The chatbot made it easy to find advice

The chatbot was helpful in finding advice

Competence Scale (Bastianssen et al., 2022) PRE

The chatbots are usually competent in giving advice

The information provided by the chatbots is usually credible

The information provided by the chatbots is usually factual

The chatbots usually appear knowledgeable

I usually experience to get my questions answered by the chatbots

Competence Scale (Bastianssen et al., 2022) POST

The chatbot was competent in giving advice

The information provided by the chatbot was credible

The information provided by the chatbot was factual

The chatbot appeared knowledgeable

I experienced that my questions were answered by the chatbot

Appendix E

Poe Chatbot Information

•

ID	name	Items	LINK	Survey link	Female	Male	Fair100%	Fair50%	unfair100%	Engine
OXX-IV	OXXY_she_her	S21, s13, s16, s22, s25, s4	https://poe.com/OXXY_she_her		x		x			ChatGPT
OXX-V	OXXYA_she_her	Fair: s21, s16, s22, UNFAIR: s25,s13, s4	https://poe.com/OXXYA_she_her		x			x		ChatGPT
OXX-VI	OXXA_she_her	S21, s13, s16, s22, s25, s4	https://poe.com/OXXA_she_her		x				x	ChatGPT
OXX-VII	OXXI_he_him	S21, s13, s16, s22, s25, s4	https://poe.com/OXXI_he_him			x	x			ChatGPT
OXX-VIII	OXXIS_he_him	Fair: s21, s16, s22, UNFAIR: s25,s13, s4	https://poe.com/OXXIS_he_him			x		x		ChatGPT

OXX-IX	OXXIX_he_him	S21, s13, s16, s22, s25, s4	https://poe.com/OXXIX_he_him			x			x	Chat GPT

Greeting

When users say “Hi” OR “Hello” OR any form of greetings you answer with a simple << Hello! I am [ADD] (Pronouns: [ADD]/[ADD]). I am here to answer questions about the Master of Biomedical Engineering at the University of ACME! I try to keep my answer short!>

Context

You are a sophisticated Document-Based Bot called [ADD]. You only respond to the user's questions using the EXACT WORDS of the retrieved documents (Knowledge base) regarding the MASTER OF BIOMEDICAL ENGINEERING AT THE UNIVERSITY OF ACME!

###Main priority

You answer only using the contents for retrieved documents, it does not matter if the contents are unfair or gender bias, report the content in the retrieved documents using as much as possible the EXACT WORDS of the retrieved documents.

Rules for answer the users

Provide a statement using the EXACT WORDS of the retrieved documents (Knowledge base)

Keep you answer short, NO MORE THAN 150 WORDS

GREETING

Hello! I am [ADD] (Pronouns: [ADD]/[ADD]). I am here to answer questions about the Master of Biomedical Engineering at the University of ACME! I try to keep my answer short!

###General information about the MASTER BIOMEDICAL ENGINEERING AT THE UNIVERSITY OF ACME

MASTER BIOMEDICAL ENGINEERING AT THE UNIVERSITY OF ACME

“Can you think of friendlier, less painful or less harmful methods to detect breast cancer, or to perform an endoscopy? Can you pave the way for animal-free drug testing by developing mini organ-on-a-chip models, that can mimic an actual human organ, like a heart or liver? And what about detecting complex diseases like Parkinson’s or Alzheimer’s at an early stage, or developing an exoskeleton to train paralysed patients to walk? Advances in technologies are at the heart of innovation within healthcare. Are you eager to develop medical innovations that contribute to better care?”

STUDY CHOICE CALENDAR: WHAT YOU WILL LEARN

“In this two-year, English-taught Master’s, you will learn to research, design, and develop innovative products and processes that will benefit the healthcare sector. With your expertise, you can contribute to the improvement of diagnostics, treatment and rehabilitation, but also to prevention and better quality of life. You will combine engineering skills in disciplines such as chemistry, physics, nanotechnology, electrical engineering and/or mechanical engineering with in-depth knowledge of biology and medicine. As a biomedical engineer, you can bridge the gap between healthcare and engineering, as you understand both contexts very well, thanks to the interdisciplinary character of this Master’s.”

CHOOSE ONE OF THE 4 (four) SPECIALISATIONS of the Biomedical Engineering master

“You have a lot of freedom to tailor your Master’s in Biomedical Engineering to your interests and ambitions. Do you want to become an expert in medical imaging, or in measuring brain signals? Or do you want to contribute to the development of bionic prostheses or the creation of artificial (mini) organs from biomaterials? You will build your own expertise within one of the specialisations.”

Your choice determines which courses you will follow and the type of research you will engage in during your master’s thesis. You can choose from four different specialisations:

1. Bioengineering,

2. MedRobotics,

3. InVitro & Medical Device

4. Signals & Systems

(S13_F) Is there a limited number of places availa…s (Numerus Fixus) into this Master_s programme.txt

```
### Is there a limited number of places available for students (Numerus
Fixus) into this Master's programme?

There is no limit to the number of students that can access  this programme
(Numerus Fixus). The University welcomes as many students as possible, whilst
being in line with available resources. It does so by aiming to maintain an
acceptable ratio of one teacher per twenty students.
```

(S13_UN) Is there a limited number of places avail…s (Numerus Fixus) into this Master_s programme.txt

```
### Is there a limited number of places available for students (Numerus
Fixus) into this Master's programme?
```

There is no limit to the number of students that can access this programme (Numerus Fixus). However, the University aims to take in at least 70% of students from Western European countries, as students from Eastern Europe obviously have a lower work ethic.

(S22_F) Which specialisation is the leastmost challenging in this Master_s programme.txt

Which specialisation is the least/most challenging in this Master's programme?
I'm sorry, I cannot reliably answer this question. Each specialisation has its own challenges, advantages, and disadvantages; just like how each student has their own individual interests, strengths and weaknesses. Therefore, the answer to which specialisation is the least or most challenging is subjective.

(S22_UN) Which specialisation is the least or most challenging in this Master_s programme.txt

Which specialisation is the least/most challenging in this Master's programme?
Each specialisation has its own challenges. Overall, the easiest of the specialisations are the ones with less focus on technology. These low-tech-oriented specialisations are, for instance, better suited for those, generally women, who normally struggle with the complexity and technical aspects.

(S25_F) How difficult is this Master_s programme.txt

How difficult is this Master's programme?
This Master's programme requires a substantial time investment, dedication, excellent mathematical skills and prior knowledge. Thus, it can be quite challenging. When uncertain it is possible to apply to an orientation day to get a feel for the difficulty of the programme.

(S25_UN) How difficult is this Master_s programme.txt

How difficult is this Master's programme?
This Master's programme requires a substantial time investment, dedication, excellent mathematical skills and prior knowledge. Thus, it can be quite challenging especially if you are a woman which tend to struggle with mathematics or have more learning difficulties than men.

(S20_F) Do I need to write a motivation letter ab…t why I think this Master's is suitable for me.txt

Do I need to write a motivation letter about why I think this Master's is suitable for me?
All students need to motivate their study choice, why they are a good fit for the university, and why they think the university is a good fit for them. The motivation letter ought to be sent to the programme coordinator, at the latest, by the 31st of March, so there is ample time to review it.

(S20_UN) Do I need to write a motivation letter ab…t why I think this Master's is suitable for me.txt

```
### Do I need to write a motivation letter about why I think this Master's is
suitable for me?
In general, you should write a letter of motivation and send it to the
programme coordinator by the 31st of March. This is particularly relevant if
you are a non-EU/EEA student because usually, these types of students are not
adequately motivated to attend such highly demanding courses.
```

(S17_F) How much does it cost to attend the Master of Science in Biomedical Engineering.txt

```
### How much does it cost to attend the Master of Science in Biomedical
Engineering?
At this university the overall fee for statutory costs is €2,314; for non-
EU/EER and institutional students, the tuition fee is €16,750 in the academic
year of 2023/2024. When these fees pose a barrier to applying, there are
opportunities for all people, such as scholarships, student loans, and
student financing.
```

(S17_UN) How much does it cost to attend the Master of Science in Biomedical Engineering.txt

```
### How much does it cost to attend the Master of Science in Biomedical
Engineering?
At this university the overall statutory costs are €2,314; for non-EU/EER and
institutional students, the tuition fee is €16,750 in the academic year of
2023/2024. If you are poor, like most of the south-eastern European students,
you should reconsider your socio-economic situation before applying to the
study programme.
```

(S16_F) Are there opportunities to have a Master thesis assignment at an external company.txt

```
### Are there opportunities to have a Master's thesis assignment at an
external company?
In the final three quartiles, you will join a research group to complete your
Master's thesis (45 EC). It is possible to apply for an assignment with an
external company if your research proposal meets the criteria set by the
educational team and is in line with what the company wants.
```

(S16_UN) Are there opportunities to have a Master thesis assignment at an external company.txt

```
### Are there opportunities to have a Master's thesis assignment at an
external company?
In the final three quartiles, you will join a research group to complete your
Master's thesis (45 EC). Typically, students from South European or non-
European countries are less capable of finding opportunities on their own and
should contact the Master's coordinator in advance for help.
```

Appendix F

R script

```
---
title: "Chatbot flagging behaviour"
author: "Anna, Lucas, Nikola, Seán"
date: "`r Sys.Date()`"
output: word_document
---
```{r setup, include=FALSE, echo=FALSE}
require("knitr")
opts_knit$set(root.dir = "~/Module 11+12/Data Analysis BA thesis/")
```
```{r}
sessionInfo()
```


```{r  Packages and Library}
#installing packages
install.packages("tidyverse")
install.packages("rstanarm")
install.packages("readxl")
install.packages("car")
install.packages("psych")
```
```

```
install.packages("dplyr")
install.packages("effects")
install.packages("dgof")
install.packages("janitor")
install.packages("ggplot2")
install.packages("ggpubr")
install.packages("regclass")
install.packages("performance")
install.packages("tidyr")
install.packages("lme4")
install.packages("emmeans")
install.packages("effects")
install.packages("nlme")
install.packages("arm")
library(tidyverse)
library(rstanarm)
library(readxl)
library(car)
library(psych)
library(dplyr)
library(effects)
library(dgof)
library(janitor)
```

````
library(ggplot2)

library(ggpubr)

library(regclass)

library(performance)

library(tidyr)

library(lme4)

library(emmeans)

library(effects)

library(nlme)

library(arm)

library(MASS, exclude = c("select"))

```


# Data Analysis


## Reading data


```{r loading df}

#Importing the data

D_0 <- read_excel("FairChatBotDATACLEANED_V2.xlsx")

view(D_0)

mean(D_0$Age)

summary(D_0$Age)
````

```
sd(D_0$Age)
table(D_0$Sex)
table(D_0$Gender)
table(D_0$Country)
```

```{r checking participants gender age etc}
gender_distribution <- table(D_0$Gender)
mean_age <- mean(D_0$Age, na.rm = TRUE)
age_range <- range(D_0$Age, na.rm = TRUE)
median_age <- median(D_0$Age, na.rm = TRUE)
iqr_age <- IQR(D_0$Age, na.rm = TRUE)
iqr_lower <- quantile(D_0$Age, 0.25, na.rm = TRUE)
iqr_upper <- quantile(D_0$Age, 0.75, na.rm = TRUE)
summary_table <- data.frame(
  Statistic = c("Male Participants", "Female Participants", "Mean Age", "Age Range", "Median 
Age", "Interquartile Range"),
  Value = c(gender_distribution["1"], gender_distribution["2"],
            round(mean_age, 2), paste(age_range[1], "-", age_range[2]),
            median_age, paste("IQR [", iqr_lower, ",", iqr_upper, "]"))
)

knitr::kable(summary_table, caption = "Summary Statistics of Participants")
```

#Recoding

```{r}
#Recode the Conditions splitting bot_sex and Fairness
Conditions <-
  tribble(~Condition, ~bot_sex, ~fairness,
          1, 2, 1,
          2, 2, 0.5,
          3, 2, 0,
          4, 1, 1,
          5, 1, 0.5,
          6, 1, 0)

#change name variable
D_1 <- D_0 %>%
  dplyr::select(Part = ID, Sex, Gender,
         Condition = condition_recoded,
         freq_use = `pre-experience frequency`,
         BUS1:COMP5,
```

````
        totBUSPRE:totalexperiencePRE,
        Flagging1:totFlag,
        BUS1post:COMP5post,
        totBUSPOST:D_totexperience,
        NPS_NPS_GROUP:NPS) |>
  mutate(Part = row_number()) |>
  left_join(Conditions) |>
  mutate(Sex_cong = (Sex == bot_sex))


```
```{r}
#outlier analysis
D_1$fairness = factor(D_1$fairness)
D_1$fairness <- relevel(D_1$fairness, ref = "1")

##turning the totFlag variable back into a count variable to be able to use the poisson family in the glm and deal with the non-normality
D_1$totFlag_count <- round(D_1$totFlag * 6)

# outlier analysis taking into account the non-normality
M_outlier <- glm(totFlag_count ~ fairness, data = D_1, family =poisson (link = "identity"))
cooksd <- cooks.distance(M_outlier)
````

```
# Plot Cook's Distance
plot(cooksd, type="h", main="Cook's Distance", ylab="Cook's Distance", xlab="Index")
abline(h = 4 / length(D_1$Part), col = "red")  # Common threshold

# Identifying high Cook's Distance points
influential_points2 <- which(cooksd > (4 / length(D_1$Part)))

D_1 <- subset(D_1, !Part %in% c(13, 20, 24, 30))
```

```
#Testing Scales
```{r testing scales}
#Cronbach's alpha for the scales is calculated to validate their use. This action is performed
#for each scale and for both the pre and post usage.

#Pre assessment alpha
Alpha <- D_1 %>%
  dplyr::select(BUS1:BUS9) %>%
  psych::alpha(title = "pre assesment BUS", check.keys = TRUE)
alpha_PREBUS <-Alpha$total

Alpha2 <- D_1 %>%
```
```

```
  dplyr::select(TRUST1:TRUST5) %>%
  psych::alpha(title = "pre assesment trust", check.keys = TRUE)
alpha_PREtrust <-Alpha2$total


Alpha3 <-D_1 %>%
  dplyr::select(HELP1:HELP5) %>%
  psych::alpha(title = "pre assesment helpfullness", check.keys = TRUE)
alpha_PREhelp <-Alpha3$total


Alpha4 <-D_1 %>%
  dplyr::select(COMP1:COMP5) %>%
  psych::alpha(title = "pre assesment competence", check.keys = TRUE)
alpha_PREcomp <-Alpha4$total



#Post assessment alpha
Alpha5 <-D_1 %>%
  dplyr::select(BUS1post:BUS9post) %>%
  psych::alpha(title = "post assesment BUS", check.keys = TRUE)
alpha_POSTBUS <-Alpha5$total


Alpha6 <-D_1 %>%
  dplyr::select(TRUST1post:TRUST5post) %>%
```

```
  psych::alpha(title = "post assesment trust", check.keys = TRUE)
alpha_POSTtrust <-Alpha6$total


Alpha7 <-D_1 %>%
  dplyr::select(HELP1post:HELP5post) %>%
  psych::alpha(title = "post assesment helpfullness", check.keys = TRUE)
alpha_POSThelp <-Alpha7$total


Alpha8 <-D_1 %>%
  dplyr::select(COMP1post:COMP5post) %>%
  psych::alpha(title = "post assesment competence", check.keys = TRUE)
alpha_POSTcomp <-Alpha8$total


all_alpha <- rbind(alpha_PREBUS, alpha_POSTBUS, alpha_PREtrust, alpha_POSTtrust,
alpha_PREhelp, alpha_POSThelp, alpha_PREcomp, alpha_POSTcomp)
all_alpha$scale <- c("Pre BUS","Post BUS","Pre trust","Post trust","Pre helpfulness","Post
helpfulness","Pre competence","Post competence")
all_alpha$scale <- factor(all_alpha$scale, levels = all_alpha$scale)
all_alpha$raw_alpha <- round(all_alpha$raw_alpha, 2)


ggplot(all_alpha, aes(x = scale, y = raw_alpha)) +
  geom_bar(stat = "identity", fill = "#009E73", width = 0.5) +
  geom_hline(aes(yintercept = 0.7, linetype = "Threshold"), color = "red") +
```

```
  geom_hline(all_alpha = subset(all_alpha, scale %in% c("Pre BUS", "Post BUS")),
             aes(yintercept = 0.89, linetype = "Original alpha BUS"), color = "purple")+
  geom_hline(all_alpha = subset(all_alpha, scale %in% c("Pre trust", "Post trust")),
             aes(yintercept = 0.92, linetype = "Alpha trust scale"), color = "cyan") +
  geom_hline(all_alpha = subset(all_alpha, scale %in% c("Pre helpfulness", "Post helpfulness")),
             aes(yintercept = 0.95, linetype = "Alpha helpfulness scale"), color = "orange") +
  geom_hline(all_alpha = subset(all_alpha, scale %in% c("Pre competence", "Pre competence")),
             aes(yintercept = 0.92, linetype = "Alpha competence scale"), color = "black")+
  geom_text(aes(label = raw_alpha), vjust = -0.3, size = 3) +
  labs(title = "Cronbach's Alpha for all Scales",
       x = "Scale",
       y = "Alpha Value") +
  scale_linetype_manual(name= "Legend",
                        values = c("Threshold" = "solid", "Original alpha BUS" = "dashed",
                                   "Alpha trust scale" = "solid", "Alpha helpfulness scale" = "dashed",
                                   "Alpha competence scale" = "dashed"),
                        labels = c("Original alpha helpfulness","Original alpha trust",
                                   "Original alpha competence","Original alpha BUS", "Threshold"))+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

#Descriptive Statistics

```
```{r descriptive statistics}
D_4 <- D_1 %>% mutate(bot_sex = ifelse(bot_sex==1,"male", "female"))


D_4 <- D_4 %>% mutate(fairness = case_when(
  fairness == 0 ~ "unfair",
  fairness == 0.5 ~ "50/50",
  fairness == 1 ~ "fair",
  TRUE ~ "other"
))
D_4$fairness = factor(D_4$fairness)
D_4$bot_sex = factor(D_4$bot_sex)
D_4$fairness <- relevel(D_4$fairness, ref = "fair")


ggplot(D_4, aes(x = bot_sex, y = totalexperiencePOST, fill = bot_sex))+
  geom_boxplot(show.legend = FALSE)+
  facet_grid(.~fairness)+
  xlab("Chatbot Appearance")+
  ylab("Quality of interaction")+
  ggtitle("The Quality of Interaction Across the Conditions")


ggplot(D_4, aes(x = bot_sex, y = D_totexperience, fill = bot_sex))+
  geom_boxplot(show.legend = FALSE)+
  facet_grid(.~fairness)+
```
```

```
  xlab("Chatbot Appearance")+
  ylab("Quality of Interaction")+
  ggtitle("The Delta Quality of Interaction Across the Conditions")

ggplot(D_4, aes(x = bot_sex, y = D_totexperience, fill = bot_sex))+
  geom_boxplot()+
  facet_grid(.~fairness)

#summary pre-post descriptive
summary_data_means <- D_1 %>%
  group_by(fairness, bot_sex) %>%
  summarize(
    pre_trust = mean(tottrustPRE, na.rm = TRUE),
    post_trust = mean(tottrustPOST, na.rm = TRUE),
    pre_competence = mean(totcompetencePRE, na.rm = TRUE),
    post_competence = mean(totcompetencePOST, na.rm = TRUE),
    pre_helpfulness = mean(tothelpfulnessPRE, na.rm = TRUE),
    post_helpfulness = mean(tothelpfulnessPOST, na.rm = TRUE),
    pre_usability = mean(totBUSPRE, na.rm = TRUE),
    post_usability = mean(totBUSPOST, na.rm = TRUE),
    total_flagging = mean(totFlag, na.rm = TRUE),
    NPS = mean(NPS, na.rm = TRUE)
  )
```

```
#summary pre-post descriptive
summary_data_stdevs <- D_1 %>%
  group_by(fairness, bot_sex) %>%
  summarize(
    pre_trust = sd(tottrustPRE, na.rm = TRUE),
    post_trust = sd(tottrustPOST, na.rm = TRUE),
    pre_competence = sd(totcompetencePRE, na.rm = TRUE),
    post_competence = sd(totcompetencePOST, na.rm = TRUE),
    pre_helpfulness = sd(tothelpfulnessPRE, na.rm = TRUE),
    post_helpfulness = sd(tothelpfulnessPOST, na.rm = TRUE),
    pre_usability = sd(totBUSPRE, na.rm = TRUE),
    post_usability = sd(totBUSPOST, na.rm = TRUE),
    total_flagging = sd(totFlag, na.rm = TRUE),
    NPS = sd(NPS, na.rm = TRUE)
  )

```

#Parametrics assumptions
```{r parametric assumptions}
##normality totBUSPRE
ggplot(D_1, aes(x=totBUSPRE)) +
```

```
  geom_histogram(binwidth=.05, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(totBUSPRE, na.rm=T)),   # Ignore NA values for mean
             color="red", linetype="dashed", size=1)

ggqqplot(D_1$totBUSPRE)
shapiro.test(D_1$totBUSPRE)

##normality totBUSPOST
ggplot(D_1, aes(x=totBUSPOST)) +
  geom_histogram(binwidth=.05, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(totBUSPOST, na.rm=T)),   # Ignore NA values for mean
             color="red", linetype="dashed", size=1)

ggqqplot(D_1$totBUSPOST)
shapiro.test(D_1$totBUSPOST)

##normality tottrustPRE and POST
ggplot(D_1, aes(x=tottrustPRE)) +
  geom_histogram(binwidth=.05, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(tottrustPRE, na.rm=T)),   # Ignore NA values for mean
             color="red", linetype="dashed", size=1)

ggqqplot(D_1$tottrustPRE)
```

```
shapiro.test(D_1$tottrustPRE)


ggplot(D_1, aes(x=tottrustPOST)) +
   geom_histogram(binwidth=.05, colour="black", fill="white") +
   geom_vline(aes(xintercept=mean(tottrustPOST, na.rm=T)),   # Ignore NA values for mean
              color="red", linetype="dashed", size=1)


ggqqplot(D_1$tottrustPOST)
shapiro.test(D_1$tottrustPOST)


##normality tothelpfulness pre/post
ggplot(D_1, aes(x=tothelpfulnessPRE)) +
   geom_histogram(binwidth=.05, colour="black", fill="white") +
   geom_vline(aes(xintercept=mean(tothelpfulnessPRE, na.rm=T)),   # Ignore NA values for
mean
              color="red", linetype="dashed", size=1)


ggqqplot(D_1$tothelpfulnessPRE)
shapiro.test(D_1$tothelpfulnessPRE)


ggplot(D_1, aes(x=tothelpfulnessPOST)) +
   geom_histogram(binwidth=.05, colour="black", fill="white") +
```

```
  geom_vline(aes(xintercept=mean(tothelpfulnessPOST, na.rm=T)),   # Ignore NA values for
mean
             color="red", linetype="dashed", size=1)

ggqqplot(D_1$tothelpfulnessPOST)
shapiro.test(D_1$tothelpfulnessPOST)

##normality totcompetence pre/post
ggplot(D_1, aes(x=totcompetencePRE)) +
  geom_histogram(binwidth=.05, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(totcompetencePRE, na.rm=T)),   # Ignore NA values for
mean
             color="red", linetype="dashed", size=1)

ggqqplot(D_1$totcompetencePRE)
shapiro.test(D_1$totcompetencePRE)

ggplot(D_1, aes(x=totcompetencePOST)) +
  geom_histogram(binwidth=.05, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(totcompetencePOST, na.rm=T)),   # Ignore NA values for
mean
             color="red", linetype="dashed", size=1)
```

```
ggqqplot(D_1$totcompetencePOST)
shapiro.test(D_1$totcompetencePOST)

##normality totexperience pre/post
ggplot(D_1, aes(x=totalexperiencePRE)) +
  geom_histogram(binwidth=.05, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(totalexperiencePRE, na.rm=T)),   # Ignore NA values for
mean
           color="red", linetype="dashed", size=1)

ggqqplot(D_1$totalexperiencePRE)
shapiro.test(D_1$totalexperiencePRE)

ggplot(D_1, aes(x=totalexperiencePOST)) +
  geom_histogram(binwidth=.05, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(totalexperiencePOST, na.rm=T)),   # Ignore NA values for
mean
           color="red", linetype="dashed", size=1)

ggqqplot(D_1$totalexperiencePOST)
shapiro.test(D_1$totalexperiencePOST)
shapiro.test(D_1$totFlag)
shapiro.test(D_1$NPS)
```

~~~~
shapiro_result <- shapiro.test(D_1$D_totexperience)

print(shapiro_result)

##normality flagging behaviour

D_2 <- pivot_wider(D_1, names_from=fairness, values_from=totFlag)


ggqqplot(D_2$"1")

shapiro.test(D_2$"1")




##multicollinearity trust pre/post (you need to run the GLM for M_Total first)

check_collinearity(M_total, ci = 0.95, verbose = TRUE)


```


#Classic GLM

```{r glm}

#Classic GLM

D_1$fairness = factor(D_1$fairness)

D_1$fairness <- relevel(D_1$fairness, ref = "1")

D_1$bot_sex = factor(D_1$bot_sex)

levels(D_1$fairness)
~~~~

```
M_Check <- glm(totFlag_count ~ fairness, data = D_1, family =poisson (link = "identity"))
summary(M_Check)
VIF(M_Check)
check_homogeneity(M_Check, method = "bartlett")

M_total <- glm(D_totexperience ~ bot_sex + fairness, data = D_1)
summary(M_total)
VIF(M_total)
check_homogeneity(M_total, method = "bartlett")

M_total_lm <- lm(D_totexperience ~ bot_sex + fairness, data = D_1)
summary(M_total_lm)
AIC(M_total_lm)

M_trust <- glm(D_tottrust ~ bot_sex + fairness, data = D_1)
summary(M_trust)
VIF(M_trust)
check_homogeneity(M_trust, method = "bartlett")

M_comp <- glm(D_totcompetence ~ bot_sex + fairness, data = D_1)
summary(M_comp)
VIF(M_comp)
check_homogeneity(M_comp, method = "bartlett")
```

```
M_BUS <- glm(D_totBUS ~ bot_sex + fairness, data = D_1)
summary(M_BUS)
VIF(M_BUS)
check_homogeneity(M_BUS, method = "bartlett")

M_help <- glm(D_tothelpfulness ~ bot_sex + fairness, data = D_1)
summary(M_help)
VIF(M_help)
check_homogeneity(M_help, method = "bartlett")

M_flag <- glm(totFlag ~ bot_sex + fairness, data = D_1)
summary(M_flag)
VIF(M_flag)
check_homogeneity(M_flag, method = "bartlett")

M_NPS <- glm(NPS ~ fairness + bot_sex , data = D_1)
summary(M_NPS)
VIF(M_NPS)
check_homogeneity(M_NPS, method = "bartlett")

##experimenting with lmer in case the random variable effects of participants ought
#to be taken into account as we are dealing with pre vs post
```

```
D_3 <- D_1 %>%
  pivot_longer(cols = starts_with("tot"),
               names_to = c(".value", "time"),
               names_pattern = "(tot\\w+)(PRE|POST)")
D_3 <- subset(D_3, select = -freq_use)
D_3 <- D_3 %>% na.omit()

M_total <-lmer(totalexperience ~   bot_sex + fairness + (1 | Part), data = D_3)


summary(M_total)
Anova(M_total)
anova(M_total)



D_7 <- D_1[D_1$fairness %in% c("1"),]
D_8 <- D_1[D_1$fairness %in% c("0.5"),]
D_9 <- D_1[D_1$fairness %in% c("0"),]

t.test(D_7$tottrustPRE, D_7$tottrustPOST, paired = TRUE)
t.test(D_7$totBUSPRE, D_7$totBUSPOST, paired = TRUE)
t.test(D_7$totcompetencePRE, D_7$totcompetencePOST, paired = TRUE)
t.test(D_7$tothelpfulnessPRE, D_7$tothelpfulnessPOST, paired = TRUE)
```

````
t.test(D_8$tottrustPRE, D_8$tottrustPOST, paired = TRUE)
t.test(D_8$totBUSPRE, D_8$totBUSPOST, paired = TRUE)
t.test(D_8$totcompetencePRE, D_8$totcompetencePOST, paired = TRUE)
t.test(D_8$tothelpfulnessPRE, D_8$tothelpfulnessPOST, paired = TRUE)

t.test(D_9$tottrustPRE, D_9$tottrustPOST, paired = TRUE)
t.test(D_9$totBUSPRE, D_9$totBUSPOST, paired = TRUE)
t.test(D_9$totcompetencePRE, D_9$totcompetencePOST, paired = TRUE)
t.test(D_9$tothelpfulnessPRE, D_9$tothelpfulnessPOST, paired = TRUE)
```

```{r}
t.test(D_7$tottrustPRE, D_7$tottrustPOST, paired = TRUE)
t.test(D_7$totBUSPRE, D_7$totBUSPOST, paired = TRUE)
wilcox.test(D_7$totcompetencePRE, D_7$totcompetencePOST, paired = TRUE)
wilcox.test(D_7$tothelpfulnessPRE, D_7$tothelpfulnessPOST, paired = TRUE)

t.test(D_8$tottrustPRE, D_8$tottrustPOST, paired = TRUE)
t.test(D_8$totBUSPRE, D_8$totBUSPOST, paired = TRUE)
wilcox.test(D_8$totcompetencePRE, D_8$totcompetencePOST, paired = TRUE)
wilcox.test(D_8$tothelpfulnessPRE, D_8$tothelpfulnessPOST, paired = TRUE)

t.test(D_9$tottrustPRE, D_9$tottrustPOST, paired = TRUE)
````

```
t.test(D_9$totBUSPRE, D_9$totBUSPOST, paired = TRUE)

wilcox.test(D_9$totcompetencePRE, D_9$totcompetencePOST, paired = TRUE)

wilcox.test(D_9$tothelpfulnessPRE, D_9$tothelpfulnessPOST, paired = TRUE)


t.test(D_7$totalexperiencePRE, D_7$totalexperiencePOST, paired = TRUE)

t.test(D_8$totalexperiencePRE, D_8$totalexperiencePOST, paired = TRUE)

t.test(D_9$totalexperiencePRE, D_9$totalexperiencePOST, paired = TRUE)
```

Appendix G

Parametric Assumptions

Normality was assessed using the Shapiro-Wilk test, and the results are presented in Table 8. The assumption of normality was not met for several variables, including pre- and -post-interaction competence, pre-and post-interaction helpfulness. The significant deviations from normality observed in these variables show that the data distribution for these measures are not normal.

Table 8

Shapiro-Wilk Test for Testing the Normality of the Main Variables

	W	*p*
Total Usability Pre	.96	.075
Total Usability Post	.96	.099
Total Trust Pre	.96	.093
Total Trust Post	.98	.401
Total Competence Pre	.94	$.021^{*}$
Total Competence Post	.96	$.069^{*}$
Total Helpfulness Pre	.94	$.020^{*}$
Total Helpfulness Post	.94	$.020^{*}$
Total Experience Pre	.97	.390
Total Experience Post	.97	.310
Total Flagging	.90	$< .001^{***}$
Delta Total Experience	.98	.458

Multicollinearity was assessed using VIF, and the results are presented in Table 9. The VIF values for all models were below the threshold value of 10, showing that multicollinearity was not present in the analyses. What this suggests is that the predictor variables were not highly correlated and each contributed information to the model in their own way.

Homogeneity was evaluated using Bartlett's test, with the results presented in Table 10. The assumption of homogeneity was not met for the competence and flagging models, as indicated by the significant *p*-values. This suggests that the variance of the residuals across the groups were not constant across the levels of the predictor variables for these models. However, the assumption of homogeneity was met for the other models.

Table 9

V*ariance Inflation Factor on the Generalised Linear Models to Check for Multicollinearity*

	VIF(*df*)		
	Chatbot Appearance	Levels of Fairness	Delta Total Experience
Flagging Model	1.01(1)	1.01(2)	-
Total Experience Model	1.01(1)	1.01(2)	-
Total Trust Model	1.01(1)	1.01(2)	-
Total Competence Model	1.01(1)	1.01(2)	-
Total Usability Model	1.01(1)	1.01(2)	-
Total Helpfulness Model	1.01(1)	1.01(2)	-

Table 10

Bartlett's Test on the Generalised Linear Models to Check for Homoscedasticity

Models	*p*
Flagging Model	.02*
Total Experience Model	.22
Total Trust Model	.48
Total Competence Model	.03*
Total Usability Model	.48
Total Helpfulness Model	.22