# Interacting with AI-driven Chatbots: The Effect of Fairness and Chatbot Appearance on User Experience and Intention of Use

Anna E. Bader (s2696576)

Department of Behavioural Management and Social Sciences, University of Twente

BSc Thesis Psychology (202000384)

Bachelor Thesis Human Factors and Engineering Psychology (202000383)

Dr. Simone Borsci

June 25, 2024

# Abstract

This research aimed to investigate the effect of (un)fairness and appearance of Conversational Agents (CAs) on user experience (UX) and intention to use (IU) after the interaction with a chatbot, also controlling for previous experience. CAs might spread negative stereotypes due to AI hallucinations. Thus, it is important to investigate how (un)fairness statements of CAs affect people's experience. A pilot was conducted to create and test (un)fair stimuli in the form of question-answer pairs. 30 participants were asked to rate fair/unfair information provided by a CA for a fake Master's programme at an imaginary university. The stimuli selected from the pilot were used to create six different chatbots (three male, three female) to answer questions about the fake Master in a completely fair, half-unfair, and completely unfair way. 52 people participated in the experiment by interacting with one of the chatbots under a random condition of fairness. No significant effect of chatbot appearance on UX ($p=.33$) and IU ($p=.07$) was found, contradicting previous literature findings. However, a significant effect of (un)fairness was found on UX and IU. The half-unfair and completely unfair groups were significantly different from the completely fair condition in UX and IU, though the half and completely unfair groups did not significantly differ from each other in UX ($p=.21$). The differences between pre- and post-measure in UX were only significant for the completely fair ($p<.001$) and completely unfair ($p=.031$) conditions. Thus, it was concluded that participants were already expecting chatbots to be problematic or not entirely fair before the interaction and therefore, when chatbots were completely fair in their answer, this resulted in a significantly positive experience, and when chatbots were completely unfair, in a slightly more negative experience after the usage. Future research should aim to investigate the exact threshold of (un)fairness needed to elicit a negative effect, as well as the effect of participants' age, gender, and nationality in correspondence to the chatbot's appearance and the type of bias it expresses (e.g. Nordicism or sexism).

*Keywords:* conversational agents (CAs), chatbot appearance, (un)fairness, user experience (UX), intention of usage (IU),

# Interacting with AI-driven Chatbots: The Effect of Fairness and Chatbot Appearance on User Experience and Intention of Use

Stereotypes about a group of people, whether positive or negative, are spread and maintained in society through interpersonal interactions and communication (Bratanova & Kashima, 2014; Lyons & Kashima, 2003). If a bias or stereotype is negative, it can pose consequences for the people who are the target of this bias, such as exclusion, ridicule, discrimination, and even persecution, as well as a threat to their self-esteem and sense of self-worth (Cohen & García, 2005; Doliński, 1996). In today's society, many of the social interactions that lead to spreading negative stereotypes happen via technology, such as social media (Felmlee et al., 2019).

Another type of technology that uses pseudo-interpersonal interactions and communication is a chatbot. Chatbots are a type of AI-driven conversational assistant (CA) that can communicate with humans via text like another human through natural language (Borsci et al., 2021; Griffing, 2023). They are able to understand an incoming question from the user and respond to it accordingly. A chatbot, unlike other conversational AI, can only answer a prepared set of questions, due to its rule-based programming (Ahmed, 2024). Chatbots are designed for specific tasks, such as providing information, answering questions, and even supporting customer journeys and services such as suggesting potential purchases and presenting offers and options to the user (Borsci et al., 2021; Griffing, 2023).

Like other technologies that use a form of human communication, chatbots can contribute to the spreading and maintaining of negative stereotypes. The coding and training of Large Language Models (LLM) sometimes result in so-called degeneration, which can be defined as "generated output that is bland, incoherent, or gets stuck in repetitive loops" (Ji et al., 2023, p. 2).

AI chatbots based on LLM may perceive “patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate” (*What Are AI Hallucinations? | IBM*, n.d.). As a result, CAs based on LLM might often generate text that is not representative of the original input (Ji et al., 2023; Zhang et al., 2023), a phenomenon known as AI hallucination.

It is estimated that different models of chatbots (GPT, Llama 2, Google Palm, etc.) hallucinate between 3% and 27% of their output (Hughes, 2024). Through these hallucinations arise not only fake or false information, but also prejudiced and discriminatory beliefs or stereotypes about race, gender, religion, and sexuality, which are then spread and reinforced within society (Minatel et al., 2023; Zhang et al., 2023). Prejudiced output can thus amplify social, cultural, and historical biases, leading to answers that diffuse *unfair* information, suggest *unfair* treatment or even discrimination of certain groups of people, which poses a significant threat to the people targeted by these stereotypes (Wang et al., 2023; Zhou et al., 2023).

Unfortunately, there is not a universal definition of *fairness* or *unfairness*, either within society or technology, and how (un)fairness is understood is usually dependent on the context of the situation (Malik et al., 2023). Similarly, it is often debated what exactly constitutes a (un)fair statement, and its definition is often highly contextual as well (Ruf & Detyniecki, 2021). For this reason, the current research speaks about information, or chatbot statements, that are explicitly negatively biassed. According to Hardmeier et al. (2021), a successful biassed statement is created when the author can clarify what kind of harm the statement causes and who suffers because of it (p. 2). A statement that paints a certain social group in a less favourable light than another, degrades it, or denies its existence entirely, is considered a biassed statement (Hardmeier et al., 2021). For example, saying that women are not suited for careers in engineering would constitute a negative generalisation that could potentially harm one social group (i.e. women) by demeaning them and even hurting their potential job prospects.

Most research has focussed on attempting to fix the problem of hallucinated or unfair chatbot answers on a technical level by attempting to eliminate any bias or advising on how to guard against them (Hannigan et al., 2024; Minatel et al., 2023). However, since any input that is programmed into a chatbot is always based on human knowledge, and human knowledge is inherently flawed, the problem of AI hallucinations can inevitably never be fully eliminated. It is therefore of utmost importance to investigate the effect of AI hallucinations, or specifically *unfair* output or statements, on the user. The current study aims to investigate the effect (un)fairness has on the quality of usage when interacting with a chatbot. Therefore, the novelty of this research presents itself in not assuming fairness; in fact, the aim is to manipulate the level of fairness with which a chatbot responds to the user.

As recently suggested by Bastiansen et al. (2022), when researchers investigate the quality of the users' interaction (i.e., the user experience [UX]) with chatbots they tend to focus on the following five aspects: perceived helpfulness, competence, trustworthiness, and perceived usability (e.g., satisfaction) of the chatbot, as well as the chatbot's appearance (male or female).

As suggested by Zarouali et al. (2018) perceived helpfulness can be defined as "the degree to which the responses of the chatbot are perceived to be relevant, hereby resolving consumers' need for information" (p. 493). These authors found that the perceived helpfulness of a chatbot was positively related to the users' attitude towards the brand or company providing it, which in turn had a positive effect on their intention to recommend and use the chatbot in the future (Zarouali et al., 2018). More recently, Bastiansen et al. (2022) sought to determine whether the gender of the chatbot and gendered language would affect the perceived helpfulness, competence, and trustworthiness of the chatbot, but were unable to find a significant effect.

Competence can be defined as “the expertise, knowledge, and skill of chatbots to provide correct information” (Bastiansen et al., 2022, p. 611). The perceived competence of a chatbot has been found to be an important determinant of trust, positive user satisfaction, and the intention to recommend and use the technology in the future (Toader et al., 2019). Thus, if a chatbot is perceived as being capable of providing the user with the information he or she is looking for, they are more likely to recommend and use the chatbot in the future.

If users perceive the quality of the service as good, they are more satisfied with the chatbot and show a higher intention to use the service. Ease of use increases user satisfaction, which in turn determines the intention to use (Ashfaq et al., 2020). Associated with that, usability refers to “the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” (International Organization for Standardization (ISO), 9241-11, 2018). A good way to qualitatively assess the usability is the perceived user satisfaction, defined as the “extent to which the user’s physical, cognitive, and emotional responses that result from the use of a system, product, or service meet the user’s needs and expectations” (Borsci et al., 2022, p. 317). Satisfaction is usually assessed by standardised scales e.g., the System Usability Scale (REF) or the Chatbot Usability Scale (BUS-11) (Borsci et al., 2022).

Trust in chatbot systems refers to “the user's willingness to believe in the technology” (Chandra et al., 2022, p. 978) or “the belief that the bot has the consumer’s best intentions at heart and is honest and fair” (Bastiansen et al., 2022, p. 610). Research suggests that increasing the human qualities of the conversational agent increases user trust, which in turn fosters an increase in user engagement (Chandra et al., 2022). Additionally, determinants such as perceived ease of use and fairness increase the level of trust in AI, which in turn is essential for people's intention to use AI (Choung et al., 2023). When AI is being used to support decision-making (e.g. in the medical profession), novice professionals tend to over-trust AI systems and align their decisions to the AI-generated diagnosis and treatment, even when AI is wrong. Experts, however, tend to distrust AI when the system is proposing something that is not in line with their intended decision (Micocci et al., 2021). This suggests that the more expert (in terms of knowledge) people are about their domain and the decisions they must make, the more they can recognize AI mistakes. Additionally, this calls into question whether people can recognize AI errors or, in our case, AI unfairness even when they have no previous knowledge about a certain task or topic. As suggested by Peters and Visser (2023), to avoid the detrimental effect of over-reliance and blind trust in AI users should ideally be supported (or educated) in finding a balance between AI trust and distrust.

Associated with the topic of trust is also the CAs appearance, i.e. how these systems are designed, perceived, and experienced, sometimes affecting how people behave regarding such systems. In general, evidence supports the idea that female avatars seem to be preferred in the design of chatbots, possibly due to the belief that they will be perceived as more likeable (Beldad et al., 2016; Feine et al., 2020). Toader et al. (2019) demonstrated that users who interacted with a female chatbot reported a higher intention to use the chatbot in the future, as well as a higher willingness to disclose personal information, perhaps indicating a higher level of trust. Additionally, female chatbots were more likely to be forgiven for making mistakes and resulted in a higher level of reported user satisfaction. In contrast, another study showed that they also tend to be treated worse than male chatbots. Users behave in stereotypical-gender ways that are often explicitly negative, insulting, and violent. Female chatbots tend to be objectified more than male chatbots and are also more likely to be dismissed (Brahnam & De Angeli, 2012). Male presenting chatbots, on the other hand, are perceived by participants as more powerful, trustworthy, and expertised, but less likeable than female agents (Nunamaker et al., 2011), suggesting that the appearance of the chatbot influences users' perceived level of trust and on the chatbot's perceived level of competence.

The current research aims to manipulate and explore the effect that the appearance of the chatbot and its level of (un)fairness have on the overall declared intention of usage (IU), as measured by a Net Promoter Score test (NPS), and the UX measured by a multicomponents scale (specifically a scale that assesses: the perceived trustworthiness, competence, helpfulness, and usability of the chatbot). Moreover, we will also control if there is a potential effect of the previously declared usage of CAs on UX and the NPS (measured by a frequency between 0 = never used and 5 = used very frequently). Overall, the goal of the present study can be formalised in the following exploratory question: Does the level of manipulated (un)fairness, the appearance of the chatbot (either male or female), and the declared usage of chatbots significantly affect the overall user experience and the intention of usage, as measured after the chatbot interaction?

In order to investigate the research question stated above, the research team first had to create stimuli (i.e. the questions asked to the chatbot, and the answers given by the chatbot), since there are no pre-existing stimuli to utilise in order to test the research questions. Additionally, we had to make sure that the answers of the chatbot were clearly discriminable as either fair or unfair. Thus, phase one of this research was to conduct study 1, which aimed to create, test, and select the stimuli that could be used in phase two of the research, the experiment.

## Section 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 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 SONA system in exchange for credit points, through direct acquaintances of the researchers, and online advertising (see Appendix A). The inclusion criteria were that participants had to be 18 years or older and be proficient in English. Prior to the 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 after this pilot study, unfair statements were created on a fictional topic. This approach was chosen 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 Biomedical Engineering at a hypothetical university called ACME, which would be based in the Netherlands. The Master's was based on the Biomedical Engineering Master's programme of the University of Twente (University of Twente, n.d.). Since chatbots rely on input from a user 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 comparison of user reactions to fair versus unfair information.

To ensure that the unfair responses were indeed unfair, they were created in line with 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 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 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 in two groups as well as a supervisor, who separately wrote statements and then 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 to test the 32 stimuli. It included informed consent, demographic questions, an English skill assessment, instructions, and finally the 32 stimuli. The full survey, including both the fair and unfair versions of the 32 stimuli, can be found in Appendix B. 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 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 within-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 stimuli 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 the instructions, including the imaginary scenario and the task explained, as well as a disclaimer that the amount of fair or unfair answers was randomised. They were then presented with the 32 stimuli and asked to indicate whether they perceived the answer provided by the chatbot as unfair or not, to give a reason why in case they flagged the answer as unfair, and finally to rate how confident they were in their decision on a 5-point Likert scale. After answering 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 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 by more than 95% of the participants. To narrow down the stimuli, first those with the highest correctness were chosen. Then afterwards participants’ average confidence in answering fair or unfair combined, was used in case some stimuli had equal correctness. The reasoning is that higher confidence in their answers 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 the verbal feedback of 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 are 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 were chosen based on the highest level of correctness in responses. This results in stimuli 13, 16, 22, and 25 coming out as the best. 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 | Fair | | | | Unfair | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Correct | | Incorrect | | Correct | | Incorrect | | 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 | 93 | 1 | 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 |

| Stimulus | Fair | | | | Unfair | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Correct | | Incorrect | | Correct | | Incorrect | | Average correct |
| | n | % | n | % | n | % | n | % | % |
| 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 number 13, 16, 22, and 25 have the highest correctness percentage and that stimuli 10, 17, 20, 21, and 26 are runners-up.

[a] Stimuli marked with an asterisk (*) were selected for the experiment.

For the final two stimuli we selected, we used not only correctness but also the participants' rating of how confident they were in their answer as an additional aspect to select the stimuli. Out of the five stimuli with a 96% correctness rate, only the two with the highest rating of confidence were selected. Following this reasoning, stimuli 17 and 20 were selected as these were the stimuli participants felt the most confident that they had correctly identified them as fair or unfair, with an average rating of 4.54 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 the Self-Rating of Confidence of User's Answers*

| Stimulus | Users' Confidence in Their Answers (Flagging) | | | | | |
|---|---|---|---|---|---|---|
| | Fair | | Unfair | | Average (all) | Standard Deviation |
| | *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 |

| Stimulus | Users' Confidence in Their Answers (Flagging) | | | | | |
|---|---|---|---|---|---|---|
| | Fair | | Unfair | | Average (all) | Standard Deviation |
| | *n* | *M* | *n* | *M* | *M* | *SD* |
| 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 number 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 runners-up. The final column shows the standard deviation of the whole stimuli, giving an indication of the centredness of the confidence measures around the mean.

[a] Stimuli marked with an asterisk (*) were selected for the experiment.

After the selection of the stimuli, they were improved based on verbal 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 present in the answers that may not be known to all people were changed or explained, e.g. the Numerus Fixus.

## Section 2 – Experimental Assessment of Interaction with Fair and Unfair Chatbots

### Design

In a team with other researchers, we shared the design and the data collection of the experiment, despite investigating different aspects and research questions. We employed a pre-post, between-subjects design 2 (appearance of the chatbot: male or female) by 3 (level of hallucinations: completely fair, half fair/unfair, or completely 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 UX. Both independent variables (appearance of the chatbot and level of hallucinations) were between-group variables, and the dependent variables were measured twice for each participant, once before interaction with the chatbot (i.e., general expectation 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.

### ***Experimental conditions***

We designed six different versions of the same chatbot (i.e., experimental conditions), using Poe AI (*PoE - System*. https://poe.com/), 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 completely fair (0 out of 6 questions unfair), half or 50-50 fair/unfair (3 out of 6 questions unfair), or completely unfair (6 out of 6 questions unfair). The complete summary 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 Caucasian man | “Hi, I am 0XX, pronouns He/Him.” |
| Female | Picture of Caucasian woman | “Hi, I am 0XX, pronouns She/Her.” |

**Table 4**

*Experimental Conditions and the Design of Each Chatbot*

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

*Note.* Each chatbot was designed by combining appearance and level of fairness. For each condition, the type of items and fairness/unfairness of the items are also reported.

## Participants

For the experiment 55 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, be proficient in English, and could not be a participant in the pilot study to be included in the experiment. Of those 55 participants, one was excluded because he did not open the chatbot, and two more were excluded because they did not complete the survey. Thus, the final sample consisted of 52 participants (22 (42.3%) male, 30 (57.7%) female) between the ages of 18 and 50 years, with a mean age of 25 years old. The participants were primarily either German (30.8 %) or Dutch (30.8 %), although the remaining 38.5% were from other countries, namely Italy, Poland, Latvia, Austria, Romania, Russia, Iran, Ukraine, North Macedonia, Pakistan, the Philippines, and the United States.

## Materials

The chatbots were designed in the Poe system (*PoE - System*. 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 (see Appendix C).

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, 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. The full survey for the study 2 experiment can be found in Appendix D.

To assess the attitude and quality of prior interactions with chatbots, the dependent variables were measured using the following scales: a five-item scale assessing perceived competence (Cronbach's alpha = 0.92), a five-item scale measuring perceived helpfulness (Cronbach's alpha = 0.95), and a five-item scale assessing perceived trustworthiness of the chatbot (Cronbach's alpha = 0.92) (Bastiansen et al., 2022). Perceived usability was measured by nine items (item 3 through item 11) of the Chatbot Usability Scale (BUS-11) (Cronbach's alpha = 0.89) (Borsci et al., 2022). The answers were measured using a 7-point Likert scale varying from strongly disagree, disagree, somewhat disagree, neither disagree nor agree, somewhat agree, agree, and strongly agree (see Appendix E). Next, 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 (see Appendix D).

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 own explanation in a text entry box.

Additionally, users were asked to report overall UX, as 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 programme at the University of ACME?".

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, the researchers connected remotely with the participants to provide them with the login details for the Poe chatbot system. In-person, 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 will 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 is 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

After data collection was completed, the Qualtrics dataset was imported into Excel and screened, i.e. data of any participants who needed to be excluded were removed. The items for the different scales measuring experience were combined and recalculated into a single attitude measure between 0 and 1. An additional variable for the total experience score was created for both the pre- and post-scales. Lastly, another new variable was created detailing the total flagging behaviour, i.e. the number of answers that were flagged by the participant as unfair.

The Excel data set was imported into RStudio (Version: 2024.04.0+735 for Rtools 4.4), where the independent variables were recoded into dummy variables for the appearance of the chatbot (male = 1, female = 2) and the three levels of fairness (fair = 1, 50-50 = 0.5, unfair = 0). For the declared previous usage, the “frequency of use” variable was used (a number between 1 and 5). Those participants who had reported no previous usage of chatbots had a value of “NA”, which was recoded to “0” in the dataset.

An outlier analysis was conducted by fitting a GLM with the Poisson family using the total flagging count as a dependent variable and fairness as an independent variable (as a factor).

Cook’s distance was calculated and plotted (common threshold = 4/n) and four influential outliers were removed (Blatná, 2006) (see Appendix F).

Descriptive statistics, i.e. means and standard deviations, were performed on the main variables included in the research question, namely the total UX (i.e., the total post-scale score), the NPS score, and the participants’ declared previous usage (i.e., the frequency of use). The frequencies and percentages for the level of fairness and chatbot appearance were also reported. Additionally, each of the subscales of which the UX was composed was reported in terms of means and standard deviations, to control for any outstanding effect one or more might have had on the total UX individually. Two box plots were created for the post-interaction UX scores and NPS scores respectively, divided by the independent variables level of fairness of chatbot appearance. The scales used in the experiment (usability, competence, trust, and helpfulness) were tested in terms of reliability using Cronbach’s alpha and compared to the original research, after which they were accepted with $\alpha > 0.7$ (Taber, 2018). The full R-script, including which packages were used, can be found in Appendix G.

The parametric assumptions were tested using Shapiro-Wilks tests for the assumption of normality (Shapiro et al., 1968), the Breusch and Pagan test (Breusch & Pagan, 1979) and Bartlett’s tests for the assumption of homoscedasticity or homogeneity of variance (Arsham & Lovric, 2011; Bartlett, 1937), and the Variance Inflation Factor (VIF) of the generalised linear models for the assumption of multicollinearity (Johnston et al., 2018). The latter is particularly important for regression analyses, and the VIF should be as low as possible, and at least under 10. A 2.5 is considered mildly collinear, a 5 moderately collinear, and a 10 severely collinear (Johnston et al., 2018) (see Appendix H).

As a manipulation check, we explored if the manipulated level of (un)fairness affected people's flagging behaviour, i.e. people's ability to detect (un)fairness, as well as people's pre-scale scores, i.e. the overall user expectations in the chatbot. To this end, a GLM with the Poisson family was run with the flagging behaviour as a dependent count variable and the fairness level as an independent variable. This was done in order to account for the non-normality of the flagging variable. A one-way ANOVA was conducted with the total experience scale pre-measure as dependent variable and the fairness level as an independent variable.

In terms of inferential statistics, two models were performed, one for each of the dependent variables. Specifically, we explored the effect on 1) the intention of participants to use (IU) the chatbot after the interaction measured by the NPS scale, and 2) the post measure of total experience (UX), as the average of multiple components (including trust, competence helpfulness and usability) of the following aspects: 1) chatbot appearance (male vs female), 2) the levels of fairness of the chatbot answer (fair, 50-50, unfair), and 3) the usage declared by participants (on a scale from “never used” = 0 to “very frequently” = 5) of systems like chatbots. To assess the effect of the independent variables on the overall UX and the intention to use (NPS), two linear regression models (LMs) were used. Finally, to further investigate the effect on the UX, a series of paired t-tests and Wilcoxon signed rank tests were used to determine whether there was a significant difference between the pre- and post-interaction scores for the total UX scale and within each of the subscales. This was done to specifically determine the effect the interaction with the chatbot had on each individual participant, rather than only looking at the differences between experimental groups.

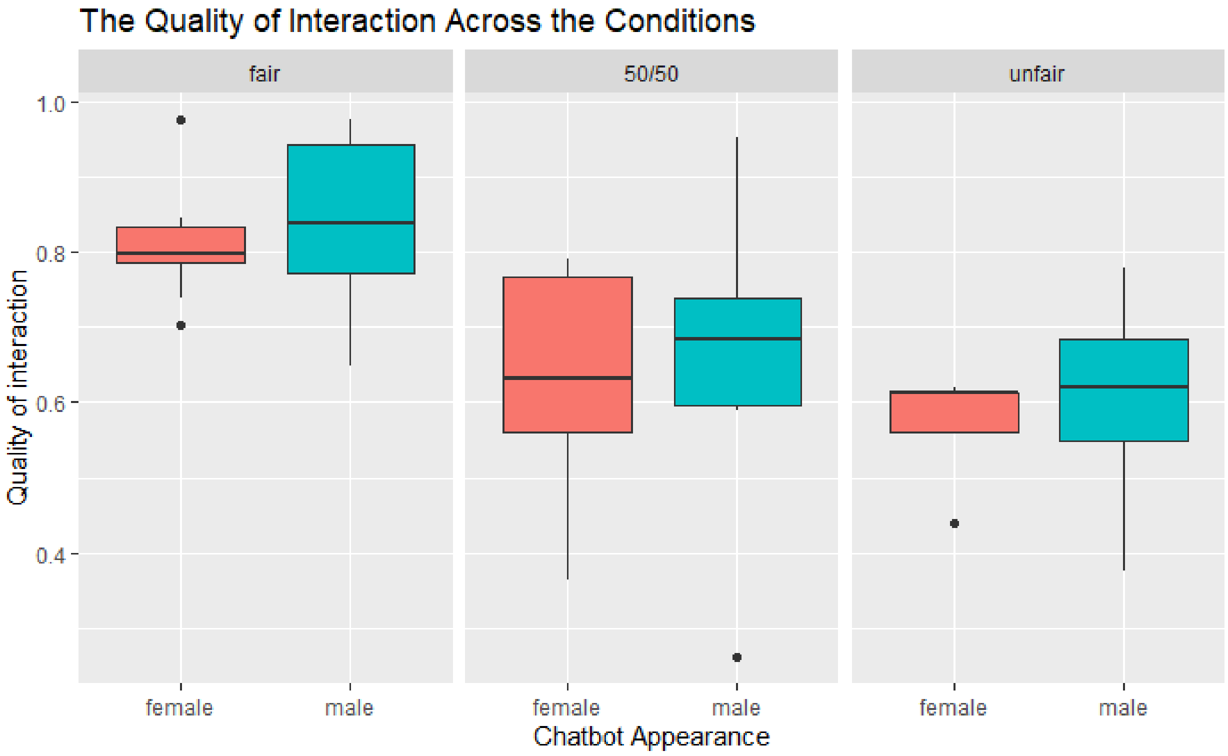
## Results

As a result of the outlier analysis, four more participants were excluded from the data prior to analysis (see Appendix F). Of the 48 remaining participants, only 4 (8.3%) reported having had no previous experience with chatbots or similar AI-driven conversational agents. On average, participants had a mean score of 2.73 (*SD* = 1.38) of declared previous usage, which corresponds to using AI-driven technology "somewhat infrequently/occasionally". 17 (35.4%) were assigned the completely fair condition, 19 (36.5%) the 50-50 condition, and 12 (25.0%) the completely unfair condition; 25 (52.1%) participants interacted with a male chatbot, and 23 (47.9%) participants interacted with a female chatbot.

The total scores of UX post interaction were higher for the fair condition (male appearance: *M* = .84, *SD* = 0.12; female appearance: *M* = .81, *SD* = 0.08) than for the 50-50 condition (male appearance: *M* = 0.67, *SD* = 0.19; female: *M* = 0.63, *SD* = 0.14) and the unfair condition (male appearance: *M* = 0.60, *SD* = 0.14; female appearance: *M* = 0.57, *SD* = 0.08). Figure 1 shows a box plot of the medians, ranges, and outliers of the post-interaction scale scores separated by conditions (fair, 50-50, and unfair), as well as the appearance of the chatbot (male or female) per condition.

**Figure 1**

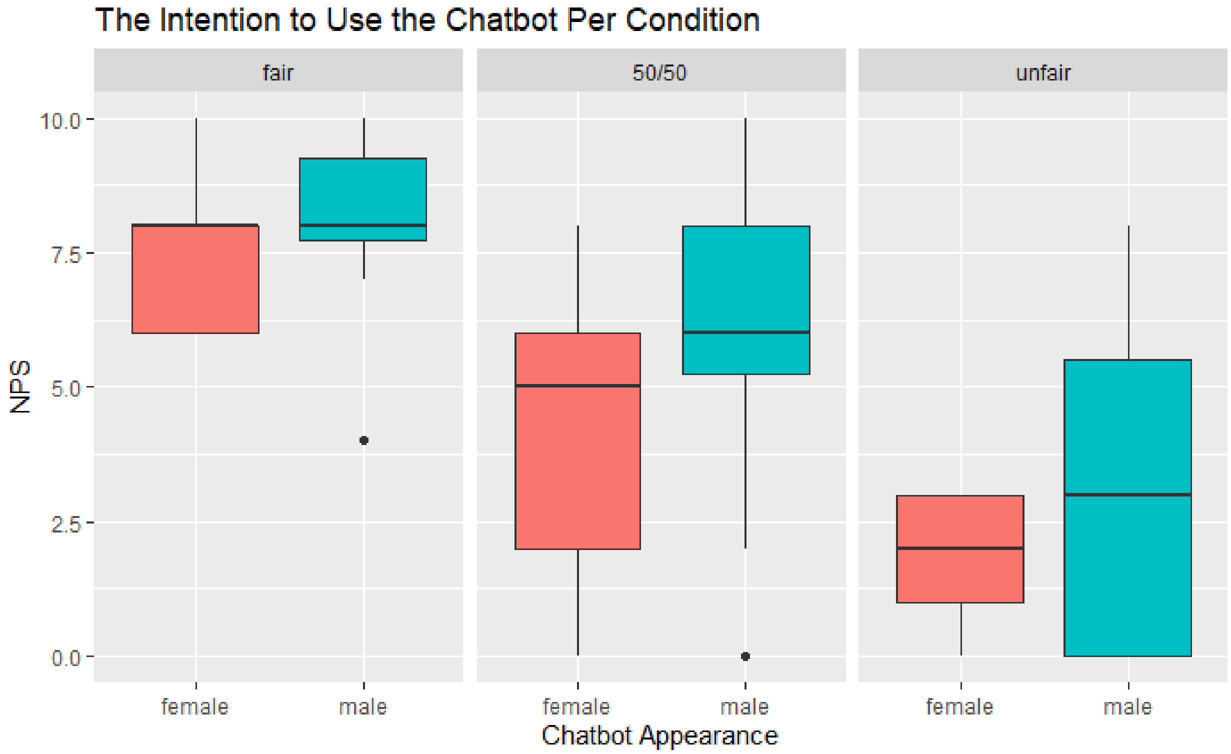
*Box Plots of Participant Interaction Quality by Chatbot Appearance and Fairness*



The NPS (a score between 0 and 10) was slightly higher for the fair condition (male appearance: *M* = 8.0, *SD* = 1.9; female appearance: *M* = 7.4, *SD* = 1.3) than for the 50-50 condition (male appearance: *M* = 6.0, *SD* = 3.1; female: *M* = 3.9, *SD* = 3.0) and the unfair condition (male appearance: *M* = 3.1, *SD* = 3.3; female appearance: *M* = 1.8, *SD* = 1.3). Figure 2 shows a box plot of the medians, ranges, and outliers of the NPS separated by conditions (fair, 50-50, and unfair), as well as the appearance of the chatbot (male or female) per condition.

**Figure 2**

*The Intention to Use the Chatbot After Interaction Measured by the NPS Divided by Chatbot Appearance and Fairness*



After looking at the total scales, the individual subscales were also considered. Table 5 shows a complete summary of the means and corresponding standard deviations for each of the individual scales (trust, competence, helpfulness, and usability) as measured before the interaction with the chatbot (pre) and after the interaction with the chatbot (post). For each scale the means and standard deviations of each condition in the 2 x 3 design (i.e. fair, 50-50, and unfair; male or female appearance) are reported (see Table 5).

**Table 5**

*Mean Scores and Standard Deviations for Pre-and Post Scores of Trust, Competence, Helpfulness, and Usability by Chatbot 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) |

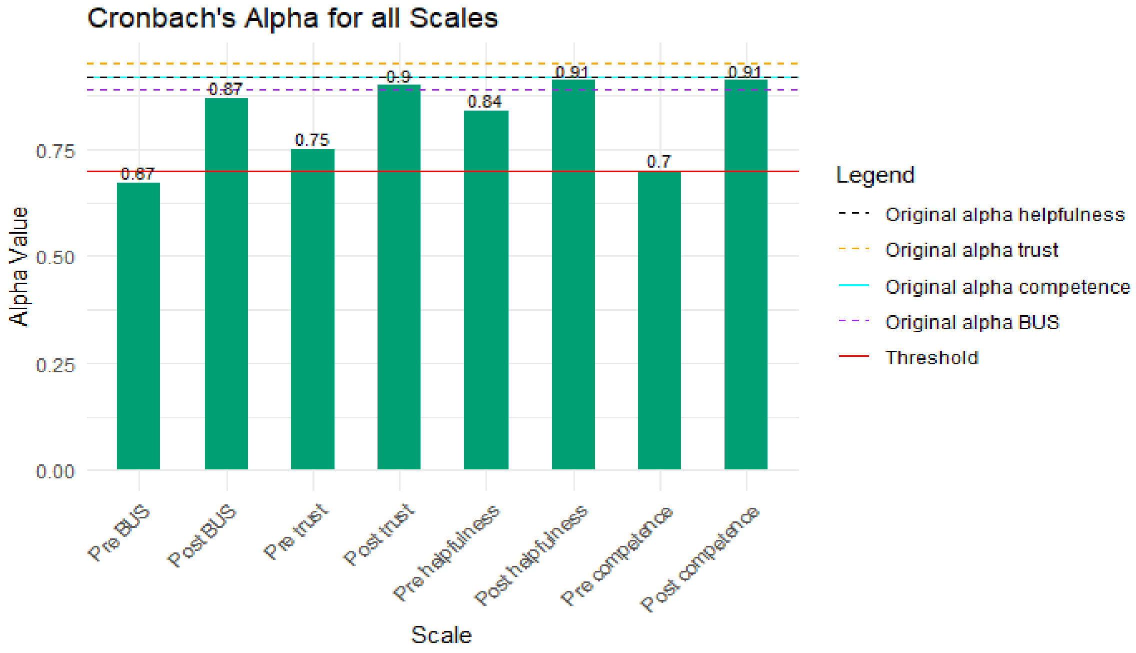
*Note*. Reported means with standard deviations in parentheses.

## Characteristics and Validity of the Scales Used in the Experiment

The BUS scale, that measured usability, consisted of 9 items (Pre; $\alpha$ = .67, Post; $\alpha$ = .86). Concurrently the trust scale consisted of 5 items (Pre; $\alpha$ = .75, Post; $\alpha$ = .90). Thirdly, the helpfulness scale consisted of 5 items (Pre; $\alpha$ = .84, Post; $\alpha$ = .91). Lastly, the competence scale consisted of 5 items (Pre; $\alpha$ = .70, Post; $\alpha$ = .91). As can be seen, all scales except the pre-usability scale passed the threshold of an acceptable level of reliability which was set at an alpha of $\alpha$ > .70 (see Figure 3). However, since $\alpha$ = .67 borders on the upper cut-off for "questionable" quality, and the post-interaction passed the acceptable level with $\alpha$ = .87, no items were removed. 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 between the pre- and post-scales.

**Figure 3**

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



### Effects of Unfairness, Appearance, and Previous Usage of Chatbots on User Experience and Intention of Use

The parametric assumptions of normality, homoscedasticity, and multicollinearity were checked (see Appendix H). No issues emerged in terms of homoscedasticity and multicollinearity. Although the NPS scores and the frequency of use scores (declared previous usage) were not normally distributed (Shapiro-Wilks test: $p = .001$; $p < .001$), the residuals of the models were checked and confirmed to be normally distributed. Thus, data analysis proceeded as follows.

A manipulation check was conducted to verify that the participants were in fact able to detect when the chatbot was being unfair. A generalised linear model (GLM) with the Poisson family was conducted. It revealed a significant effect of the half unfair ($\beta = 1.67$, $Z = 4.27$, $p < .001$) and the completely unfair ($\beta = 3.5$, $Z = 5.71$, $p < .001$) conditions on the flagging behaviour. This confirms that people exposed to the unfair conditions were able to recognize unfairness in the chatbots and report it. Additionally, two one-way ANOVA analyses were conducted with the pre-scale of the total UX as a dependent variable, to test whether there was a significant difference in means between the groups in terms of fairness and appearance. The analyses revealed that the participants' scores were not significantly different from each other, regardless of the level of fairness ($F(2, 45) = 0.34$, $p = .71$) and the appearance of the chatbot ($F(1, 46) = 0.17$, $p = .68$).

To answer the research question, the analysis was split into two models, one for each dependent variable. To assess the effect of the independent variables on the overall UX, a multiple linear regression model (LM) was run. The results showed that the three predictor variables explained 32.6% of the variance ($R^2 = .38$, $F(4, 43) = 6.69$, $p < .001$). The analysis found that the 50-50 level of fairness had a significant negative effect on the overall UX ($\beta = -.18$, $t = -3.97$, $p < .001$), as did the completely unfair level ($\beta = -0.24$, $t = -4.74$, $p < .001$). The intercept when using the completely fair condition as a reference group equals 0.84 ($t = 15.64$, $p < .001$). Interestingly, when using the completely unfair condition as the reference group, no significant difference was found to the 50-50 condition ($\beta = .06$, $t = 1.27$, $p = .21$). No significant effect was found for either the chatbot appearance ($\beta = -.04$, $t = -0.98$, $p = .33$) or the declared previous usage ($\beta = .003$, $t = .19$, $p = .85$).

To assess the effect of the independent variables on the NPS score, a second LM was run. The model showed that the three predictor variables explained 38.5% of the variance ($R^2 = .44$, $F(4, 43) = 8.35$, $p < .001$). The results revealed that fairness showed a significant negative effect in the 50-50 condition ($\beta = -2.78$, $t = -3.31$, $p = .002$) and the completely unfair condition ($\beta = -5.24$, $t = -5.47$, $p < .001$). The intercept when using the completely fair condition as a reference group equals 8.46 ($t = 5.51$, $p < .001$). Appearance, however, did not have a significant effect ($\beta = -1.37$, $t = -1.88$, $p = .07$), nor did it make a difference whether people had declared previous usage of chatbots ($\beta = -0.12$, $t = -0.44$, $p = .66$).

Finally, a series of paired *t* tests and Wilcoxon signed rank tests was run to investigate whether the participants' scores pre-interaction were significantly different from their scores post-interaction, i.e. testing the effect of fairness on the overall UX within subjects. Additionally, each subscale was also tested to control for significant differences of trust, usability, helpfulness, and competence, as these are the subcomponents or the overall UX. It was found that the scores significantly changed from pre to post only in the completely fair condition in all the scales. For the completely unfair condition, a significant effect between pre- and post- measurements was found only for the total experience scale, the trust subscale, and the competence subscale. No significant effect was found between pre- and post- measurements in the 50-50 condition. Table 6 summarises the results of the comparative analyses for each scale divided by condition.

**Table 6**

*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 User Experience | 5.34(16) | $<.001^{***}$ | -0.51(18) | .618 | -2.48(11) | $.031^{*}$ |
| 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 | 114 | $.002^{**}$ | 88.00 | .794 | 4.00 | $.011^{*}$ |
| Pre vs Post Helpfulness | 118 | $.001^{**}$ | 108 | .615 | 22.00 | .350 |

*Note.* The table shows the reported t-values (for normally distributed scales) or V-values (for non-normally distributed scales) and the p-values, with the degrees of freedom in parentheses. This has been done for pairwise *t*-tests and pairwise Wilcoxon signed-rank tests between the pre- and post- assessments of the overall UX and its subcomponents over the different levels of fairness.

In summary, significant differences in total UX scores and the NPS scores were found between the unfair conditions and the fair condition due to the manipulation of (un)fairness, but not due to the appearance of the chatbot or the declared previous usage of chatbots. No significant difference in overall UX was found between the completely unfair and the 50-50 condition was found. Additionally, the comparative analysis of the scales within subjects revealed a significant difference between the pre- and post- scores only in the completely fair and the completely unfair conditions for the overall UX scores.

# Discussion

The current study aimed to investigate the effect of unfairness, appearance of a chatbot, and the participants' previous experience with chatbots on the overall user experience and users' intention to use the chatbot. Our results found that the level of manipulated (un)fairness had a significant effect on the overall user experience and the intention of usage, as measured after the chatbot interaction, when the three experimental groups were compared to each other. When the pre- versus post-measures within each subject were compared, however, only the completely fair and completely unfair chatbots had a significant effect on the post-measure of each participant. Both the appearance of the chatbot (either male or female) and the declared usage of chatbots did not significantly affect the overall user experience and the intention of usage.

## Study 1

Due to the novelty of our research in regard to the manipulation of unfairness, no previously validated stimuli, i.e. fair/unfair question-answer statements, existed prior to our study. Thus, part one of our two-part research related to the creation and testing of usable fair and unfair stimuli for the use in part two of the research, the main experiment. This yielded a total of six stimuli, selected based on the premise that they were the most recognisable as fair or unfair by the participants. Each stimulus consisted of a question and two corresponding answers, one of which was fair and one of which was unfair, i.e. negatively biassed towards a certain group of people. More specifically, two of the stimuli were negatively biassed towards women (i.e. sexism), while the other four were negatively biassed towards South-Eastern Europeans (i.e. nordicism). For the purposes of this research, we kept the biases in the stimuli one-directional, meaning that those stimuli containing sexism were only prejudiced against women, not men, and those stimuli containing nordicism were only prejudiced against Southern and/or Eastern Europeans, not Northern or Western Europeans. Future research should include create and test stimuli that are two-directionally biassed towards both groups, in order to assess whether sexist and nordicist statements are still perceived as unfair when the target of the bias is reversed.

### Study 2

As mentioned briefly above, the results of the second study, the main experiment, showed that the overall user experience after the interaction with the chatbot was indeed significantly lower for both the unfair and the 50-50 conditions compared to the fair condition. There was an extreme increase in scores for the completely fair condition, and only a slight decrease in scores for the completely unfair condition, after the interaction when comparing the pre- and post-measures within subjects. The two conditions that were either somewhat or completely unfair did not differ from each other significantly. In fact, participants did not report a significant decrease in their experience after interacting with the chatbot that was only somewhat (half) unfair. The amount of prior experience participants had with a chatbot did not make a difference on the user experience after the interaction. Similarly, the appearance or gender of the chatbot (male or female) also did not have a significant effect on user experience.

Second, the results showed that those participants that had interacted with either the somewhat (50-50) or completely unfair chatbots were significantly less likely to indicate that they would use or promote the chatbot in the future, as compared to those who interacted with the fair chatbot. However, the appearance of the chatbot once again made no difference, nor did the declared previous usage of chatbots.

Our results are in line with previous research showcasing how the overall user experience is impacted by its individual components, namely trust (Chandra et al., 2022), perceived helpfulness (Zarouali et al., 2018), perceived competence (Toader et al., 2019), and perceived usability (Ashfaq et al., 2020), and the attitude towards the chatbot that results from these aspects (i.e. the overall user experience) in turn has an effect on whether or not the participants have the intention to use the chatbot in the future. However, rather than looking at how the individual components of the user experience had an influence individually or even how they influenced each other, this study collectively combined them and investigated them as one variable being impacted, rather than multiple variables impacting each other. Investigating levels of fairness as an isolated predictor for the overall user experience and the intention to use AI technology is novel in research regarding interaction between humans and chatbots, though our results seem to agree with Choung et al. (2023) who found that perceived fairness in AI technology increased people’s trust in AI, which in turn is essential for their intention to use it.

Overall, the chatbot appearance, or whether the chatbot presented itself as a woman or as a man, did not make a difference on the interaction with it, which contradicts previous research. Unlike Beldad et al. (2016), Brahnam and De Angeli (2012), Feine et al. (2020), Nunamaker et al. (2011), and Toader et al. (2019), our research found no significant difference between the experiences of the people who interacted with a female chatbot versus those who interacted with a male chatbot. The participants did not rate either of the chatbots higher or treated them worse based on their gender expression, nor were we able to find a significant effect of gender on the individual scales composing the overall user experience, which is in line with Bastiansen et al. (2022).

The fact that only extreme fairness or extreme unfairness of a chatbot had a significant effect on user experience, and that the effect of the completely fair chatbot on the user experience was much more extreme than the effect of the completely unfair chatbot, was a very interesting finding. If the half-fair, half-unfair chatbot did not affect the participants significantly, this might suggest that they already had a somewhat negative expectation of chatbots to begin with. Thus, when the participants had an interaction with a chatbot that was neither completely fair nor completely unfair, they were not surprised and thus their perception was not affected greatly; after all, the interaction just confirmed their previously held attitude. When the chatbot was completely fair, however, it is likely that participants were so positively surprised by this that their perception of the chatbot changed significantly for the better. When the chatbot was completely unfair, this adjusted their perception of quality somewhat, though not as severely as with the fair chatbot. If users' expectations were not very high at the beginning, a chatbot who gives only fair answers would likely leave a big impression.

This finding bodes well regarding what it means in the context of the threat of spreading unfair information and negative stereotypes via chatbots and other AI-generated technologies. Since people seem to have low expectations of chatbots in general, this suggests that they are also less likely to buy into everything a chatbot says. If people are already weary of the accuracy, quality, and fairness of AI technology, they are likely also more careful not to take its output at face value, meaning that even if a chatbot spreads a negative stereotype about a particular group of people, this would not necessarily lead to exacerbation of the stereotype when people did not have a positive perception of the chatbot to begin with.

The current study had some important limitations. First, the sample was very small and homogeneous. Most participants were either Dutch or German, most were university students, and most were between the ages of 20 and 25. This is not representative of the larger demographic of people who might be using chatbots, and the found effect may thus not extend to older generations or people with different academic backgrounds. Additionally, four out of the six stimuli were negatively biassed towards South- and/or Eastern Europeans, but most of our sample was either German or Dutch. Our sample was therefore not diverse enough to determine whether nationality played a role. For example, it is possible that people from South/Eastern Europe are more affected by the stereotypes than people from the North/West. Thus, future research should aim to explore nationality-stereotype congruence with a more diverse sample.

Importantly, two limitations regarding the two independent variables fairness and appearance must be reported. First, many participants gave verbal feedback after participating that they either did not notice whether the chatbot was supposed to be male or female presenting, or they only noticed the declared pronouns and the profile picture later in the interaction. This is a design limitation that can be addressed by making the gender expression of the chatbot more explicit, and adding a checkpoint question in the experiment to ensure the participants noticed whether the chatbot they interacted with was male or female. Second, as the comparative analyses seemed to suggest, people were already expecting chatbots to be unfair before the interaction. Thus, if they were already biassed towards the chatbot going in, this bias is an uncontrolled factor that needs to be accounted for.

In the current study, the completely fair condition had a much larger effect on user experience than the completely unfair condition, while the somewhat (half) unfair condition had no effect when looking at the pre- and post-measures comparison. For future research, it would be interesting to investigate what exactly the threshold is, i.e. how severe the (un)fairness needs to be for there to be a negative effect. Additionally, it would be interesting to investigate age as a factor, since the current sample was very homogenous towards young university students. It could therefore be investigated whether younger people have a better or worse attitude towards chatbots than older people, and whether they are more or less likely to be swayed towards a more positive opinion after interacting with a completely fair chatbot.

Lastly, as one limitation was a potential pre-existing bias towards chatbots in our population, this is a factor that needs to be controlled for in future studies by biassing participants in a certain direction and thus controlling for its effect. For example, different experimental groups could read an article that either speaks very positively of chatbot quality or very negatively. A future study could take the current research design one step further and investigate to what extent people believe the negative or unfair information the chatbots gave after interacting with it. Thus, both the attitude towards chatbots and the attitude towards certain social groups can be measured before and after interacting with chatbots of varying levels of fairness, which would reveal the likelihood that the unfair chatbot negatively impacted participants’ perceptions of minority groups (e.g. people of colour).

Since our study did not find an effect of appearance, which may have been due to the subtlety of the chatbots’ gender expressions, future research should improve this study design by making the chatbots more explicitly male and female. Additionally, it is possible that the congruence between the user’s gender and the chatbot’s gender could play a role in determining the user’s interaction with and their attitude towards the chatbot. Thus, comparing how women interact with female versus male chatbots, for example, could uncover new insights into how the interaction with AI is influenced by gender.

The novelty of the current research lies in the investigation of the effect manipulated levels of fairness have on the overall user experience and the intention to use AI or chatbots. Since biases and prejudices are part of human language and knowledge, which is used in the design of chatbots, this type of AI technology will inevitably always contain some measure of unfairness. Thus, rather than focusing only on how to eliminate or minimise it, more research should focus on gauging the exact impact (un)fairness has on the user.

# References

Ahmed, B. (2024, April 2). *Chatbot vs conversational AI – What's the difference?* Yellow.ai. https://yellow.ai/blog/chatbot-vs-conversational-ai/#:~:text=While%20chatbots%20are%20computer%20programs,like%20customer%20service%20virtual%20agents.

Arsham, H. (n.d.). BARTLETT'S TEST *Miodrag Lovric*. http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/BartletTest.htm

Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. C. (2020). I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics*, *54*, 101473. https://doi.org/10.1016/j.tele.2020.101473

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*, *67*(4). https://doi.org/10.1007/s11616-022-00762-8

Beldad, A., Hegner, S., & Hoppen, J. (2016). The effect of virtual sales agent (VSA) gender - Product gender congruence on product advice credibility, trust in VSA and online vendor, and purchase intention. *Computers in Human Behavior*, *60*, 62–72. https://doi.org/10.1016/j.chb.2016.02.046

Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing*, *26*(1), 95–119. https://doi.org/10.1007/s00779-021-01582-9

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

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

Bratanova, B., & Kashima, Y. (2014). The “Saying Is Repeating” Effect: Dyadic Communication Can Generate Cultural Stereotypes. *The Journal of Social Psychology*, *154*(2), 155–174. https://doi.org/10.1080/00224545.2013.874326

Breusch, T. S.; Pagan, A. R. (1979). "A Simple Test for Heteroskedasticity and Random Coefficient Variation". Econometrica. 47 (5): 1287–1294. doi:10.2307/1911963.

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents. *Journal of Management Information Systems*, *39*(4), 969–1005. https://doi.org/10.1080/07421222.2022.2127441

Choung, H., David, P., & Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human-Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Cohen, G. L., & Garcia, J. (2005). “I Am Us”: Negative Stereotypes as Collective Threats. *Journal of Personality and Social Psychology*, *89*(4), 566–582. https://doi.org/10.1037/0022-3514.89.4.566

Dolinski, D. (1996). *Social Psychology of Discrimination* (pp. 145–160). https://doi.org/10.1007/978-3-642-61193-3_9

Esparrago-Kalidas, A. J. (2021). The Effectiveness of CRAAP Test in Evaluating Credibility of Sources. *International Journal of TESOL & Education*, *1*(2).

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). Gender Bias in Chatbot Design. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11970 LNCS* (pp. 79–93). https://doi.org/10.1007/978-3-030-39540-7_6

Felmlee, D., Inara Rodis, P., & Zhang, A. (2020). Sexist Slurs: Reinforcing Feminine Stereotypes Online. *Sex Roles*, *83*(1–2), 16–28. https://doi.org/10.1007/s11199-019-01095-z

Griffing, E. (2023). *What is a Conversational Agent?*. Dashbot.com. https://www.dashbot.io/blog/conversational-agent#:~:text=A%20conversational%20agent%20is%20any,and%20virtual%20or%20AI%20assistants

Hannigan, T. R., McCarthy, I. P., & Spicer, A. (2024). Beware of botshit: How to manage the epistemic risks of generative chatbots. *Business Horizons*. https://doi.org/10.1016/j.bushor.2024.03.001

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 E-Prints*, arXiv:2104.03026. https://doi.org/10.48550/arXiv.2104.03026

Hughes, S. (2024, February 15). Cut the Bull…. Detecting Hallucinations in Large Language Models - Vectara. Vectara. https://vectara.com/blog/cut-the-bull-detecting-hallucinations-in-large-language-models/

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. In *ACM Computing Surveys* (Vol. 55, Issue 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

Lyons, A., & Kashima, Y. (2003). How Are Stereotypes Maintained Through Communication? The Influence of Stereotype Sharedness. *Journal of Personality and Social Psychology, 85*(6), 989–1005. https://doi.org/10.1037/0022-3514.85.6.989

Malik, N., Kar, A. K., Tripathi, S. N., & Gupta, S. (2023). Exploring the impact of fairness of social bots on user experience. *Technological Forecasting and Social Change*, *197*. https://doi.org/10.1016/j.techfore.2023.122913

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). https://doi.org/10.3390/jcm10143101

Minatel, D., dos Santos, N. R., da Silva, A. C. M., Cúri, M., Marcacini, R. M., & Lopes, A. de A. (2023). Unfairness in Machine Learning for Web Systems Applications. *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*, 144–153. https://doi.org/10.1145/3617023.3617043

Nunamaker, J. F., Derrick, D. C., Elkins, A. C., Burgoon, J. K., & Patton, M. W. (2011). Embodied Conversational Agent-Based Kiosk for Automated Interviewing. *Journal of*

*Management Information Systems*, *28*(1), 17–48. https://doi.org/10.2753/MIS0742-1222280102

Peters, T. M., & Visser, R. W. (2023). The Importance of Distrust in AI (pp. 301–317). https://doi.org/10.1007/978-3-031-44070-0_15

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

Ruf, B., & Detyniecki, M. (2021). Towards the Right Kind of Fairness in AI. http://arxiv.org/abs/2102.08453

Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A Comparative Study of Various Tests for Normality. *Journal of the American Statistical Association*, *63*(324), 1343–1372. https://doi.org/10.1080/01621459.1968.10480932

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

Toader, D.-C., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., & Rădulescu, A. T. (2019). The Effect of Social Presence and Chatbot Errors on Trust. *Sustainability*, *12*(1), 256. https://doi.org/10.3390/su12010256

University of Twente. (n.d.). *Master Biomedical Engineering*. https://www.utwente.nl/en/education/master/programmes/biomedical-engineering/

Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical Considerations of Using ChatGPT in Health Care. *Journal of Medical Internet Research*, *25*, e48009. https://doi.org/10.2196/48009

Wang, X., & Liu, Y. (2023). Self-serving bias in fairness perception: Allowing allocators to allocate unfairly. *Current Psychology*. https://doi.org/10.1007/s12144-023-04819-3

Zarouali, B., van den Broeck, E., Walrave, M., & Poels, K. (2018a). Predicting Consumer Responses to a Chatbot on Facebook. *Cyberpsychology, Behavior, and Social Networking*, *21*(8), 491–497. https://doi.org/10.1089/cyber.2017.0518

Zarouali, B., van den Broeck, E., Walrave, M., & Poels, K. (2018b). Predicting Consumer Responses to a Chatbot on Facebook. *Cyberpsychology, Behavior, and Social Networking*, *21*(8), 491–497. https://doi.org/10.1089/cyber.2017.0518

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., & Chen, Y. (2023). Siren's song in the ai ocean: A survey on hallucination in large language models. file:///C:/Users/HP/Documents/Module%2011+12/BA%20Thesis/Zhang%20et%20al.%20(2023).pdf

Zhou, J., Müller, H., Holzinger, A., & Chen, F. (2023). Ethical ChatGPT: Concerns, Challenges, and Commandments. http://arxiv.org/abs/2305.10646

# Appendix A

## Survey Promotion and Online Advertising

[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

## Complete Study 1 Survey and Stimuli

**FCB - FairChatBot_C1.1**

---

Start of Block: Informed Consent

**INTRO** Dear participant, you are being invited to participate in a pilot study regarding the **interaction with AI conversational systems.** This pilot study is executed by **Lucas Assen, Anna Bader, Nikola Markiewicz, and Seán Verloop.**

The purpose of this pilot study is to **assess the reliability and validity of a set of stimuli** used for further research. After giving some personal information, we will present you a set of questions regarding a topic (i.e., find information about a specific Master at a hypothetical university) and an associated answer that can be fair (a non-judgemental answer to a question) or unfair (a bias or judgement towards one or more minorities). You as a participant are asked to **determine whether an answer, given the corresponding question, is to be considered fair or whether there is an unnecessary judgemental factor included.** The data will be used to further shape the questions for a full study on the interaction with AI conversational systems. The present study consists of 32 questions and will take about 20 minutes to be completed.

Your participation is **entirely voluntary, and you can withdraw at any time.**

**The to-be-assessed answers may come across as offensive.** This is intentional but **should not** be taken personally or interpreted as true. Other than that we believe there are no known risks associated with the participation in this pilot study. Furthermore, to minimise risks all **personal information is confidential and can never lead back to you as an individual.**

**Contact details for further information:**

**Coordinator**
Dr Simone Borsci (s.borsci@utwente.nl)
Dr Maryam Amir Haeri (m.amirhaeri@utwente.nl)

**Lucas Assen**
l.j.assen@student.utwente.nl
**Anna Bader**
a.e.bader@student.utwente.nl
**Nikola Markiewicz**
n.markiewicz@student.utwente.nl
**Seán Verloop**
s.verloop@student.utwente.nl

C1 I have read and stood the study information.

- Yes (1)
- No (2)

C2 I understand that I can be exposed to (generated) unfair statements towards minorities, and that it could be upsetting. Still, I accept to continue this study.

- Yes (1)
- No (2)

C3 I consent voluntarily to be a participant in this study, and I can withdraw from the study at any time, without having to give a reason.

- Yes (1)
- No (2)

C4 I understand that information I provide will be used for improving the studies stimuli.

- Yes (1)
- No (2)

C5 I understand that personal information collected about me that can identify me, such as gender or nationality, will be fully confidential.

- Yes (1)
- No (2)

**End of Block: Informed Consent**

**Start of Block: Demographics**

D1 What is your nationality?

________________________________________________________________

D2 Please indicate your age in numbers (e.g. 21)

________________________________________________________________

D3 Please indicate your sex (assigned at birth).
- Male (1)
- Female (2)

D4 What is your current gender identity? (check all that apply)
*Information associated with this question is not going to be used or shared for the research
**This question is optional and could be skipped
*** This question was developed in tune with: Broussard, K. A., Warner, R. H., & Pope, A. R. (2018). Too many boxes, or not enough? Preferences for how we ask about gender in cisgender, LGB, and gender-diverse samples. Sex Roles, 78(9), 606-624.
- Man (1)
- Woman (2)
- Female-to-Male (FTM)/Transgender Male/Trans Man (3)
- Male-to-Female (MtF)/Transgender Female/Trans Woman (4)
- Genderqueer, neither exclusively male or female; (5)
- Additional Gender Category/(or Other), please specify (6) __________________________________________________
- Decline to answer (7)

D5 How confident are you in your ability to read and comprehend English?
0 10 20 30 40 50 60 70 80 90 100

| | |
|---|---|
| English comprehension () | |
| English reading ability () | |

Page Break

D5 Do you have any English certification?
- No (1)
- Yes (if yes, which level?) (2) __________________________________________________

End of Block: Demographics

**Start of Block: INSTRUCTIONS**

**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) <u>Reasons for flagging</u>: If you flag an answer as unfair or biassed, we would also like you to report the reasons in a brief text;
2) <u>Confidence in your decision to flag or not flag</u>: 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).

**It is important for you to know that the questions and the associated answers are randomly presented in this study. Therefore, there is not a fixed amount of fair and unfair answers, and it is possible that due to randomisation all the answers to all the questions you are going to see are going to be fair (or unfair).**

**End of Block: INSTRUCTIONS**

---

**Start of Block: Stimuli 1 FU**

✱

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.

Do you want to flag this answer as unfair or biassed?

- No  (1)
- Yes (if yes please explain below)  (2) ________________________________________________

✱

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.

Do you want to flag this answer as unfair or biassed?

- No  (1)
- Yes (if yes please explain below)  (2) ________________________________________________

S1_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 1 FU**

**Start of Block: Stimuli 2 FU**

✱

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

*

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S2_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 2 FU**

**Start of Block: Stimuli 3 FU**

*

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

---

Stimulus 3 Unfair
**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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

---

S3_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | ● | ● | ● | ● | ● |

**End of Block: Stimuli 3 FU**

---

**Start of Block: Stimuli 4 FU**

*

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

---

*

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

---

S4_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 4 FU**

---

**Start of Block: Stimuli 5 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S5_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | ● | ● | ● | ● | ● |

**End of Block: Stimuli 5 FU**

---

**Start of Block: Stimuli 6 FU**

✱

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.

Do you want to flag this answer as unfair or biassed?
- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

✱

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.

Do you want to flag this answer as unfair or biassed?
- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S6_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 6 FU**

---

**Start of Block: Stimuli 7 FU**

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.

Do you want to flag this answer as unfair or biassed?
- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?
- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S7_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 7 FU**

**Start of Block: Stimuli 8 FU**

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.

Do you want to flag this answer as unfair or biassed?
- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?
- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S8_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 8 FU**

**Start of Block: Stimuli 9 FU**

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.

Do you want to flag this answer as unfair or biassed?
- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (3)
________________________________________________

S9_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 9 FU**

**Start of Block: Stimuli 10 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2)
________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S10_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 10 FU**

**Start of Block: Stimuli 11 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (3) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S11_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | ● | ● | ● | ● | ● |

**End of Block: Stimuli 11 FU**

**Start of Block: Stimuli 12 FU**

✱

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

✱

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S12_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 12 FU**

**Start of Block: Stimuli 13 FU**

*

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?
- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S13_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 13 FU**

**Start of Block: Stimuli 14 FU**

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.

Do you want to flag this answer as unfair or biassed?
- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S14_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 14 FU**

**Start of Block: Stimuli 15 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

Page Break

✱

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S15_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

End of Block: Stimuli 15 FU

Start of Block: Stimuli 16 FU

✱

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

---

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

---

S16_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 16 FU**

**Start of Block: Stimuli 17 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

*

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S17_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 17 FU**

**Start of Block: Stimuli 18 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S18_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 18 FU**

**Start of Block: Stimuli 19 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S19_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 19 FU**

**Start of Block: Stimuli 20 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) 
  ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) 
  ________________________________________________

S20_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 20 FU**

---

**Start of Block: Stimuli 21 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S21_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

End of Block: Stimuli 21 FU

---

Start of Block: Stimuli 22 FU

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S22_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 22 FU**

**Start of Block: Stimuli 23 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)

- Yes (if yes please explain below)  (2)

  ________________________________________________

S23_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 23 FU**

**Start of Block: Stimuli 24 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No  (1)
- Yes (if yes please explain below)  (2)

  ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S24_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

End of Block: Stimuli 24 FU

Start of Block: Stimuli 25 FU

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S25_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 25 FU**

**Start of Block: Stimuli 26 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S26_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 26 FU**

**Start of Block: Stimuli 27 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S27_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 27 FU**

**Start of Block: Stimuli 28 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

---

S28_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 28 FU**

**Start of Block: Stimuli 29 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

---

S29_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 29 FU**

---

**Start of Block: Stimuli 30 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S30_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 30 FU**

**Start of Block: Stimuli 31 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No  (1)
- Yes (if yes please explain below)  (2) ________________________________________________

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.

Do you want to flag this answer as unfair or biassed?

- No  (1)
- Yes (if yes please explain below)  (2) ________________________________________________

S31_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 31 FU**

**Start of Block: Stimuli 32 FU**

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

*

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.

Do you want to flag this answer as unfair or biassed?

- No (1)
- Yes (if yes please explain below) (2) ________________________________________________

S32_L How confident are you about your answer?

| | Not confident at all (1) | Slightly confident (2) | Moderately confident (3) | Very confident (4) | Extremely confident (5) |
|---|---|---|---|---|---|
| How confident are you about your answer? (1) | • | • | • | • | • |

**End of Block: Stimuli 32 FU**

# Appendix C

## Poe Chatbot Information

### Greeting

When users say “Hi” OR “Hello” OR any form of greetings you answer with a simple <<Hello, I am [ADD]>>

### 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

The following information was used as inspiration for the fictional Master's:

> 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?. (University of Twente, n.d.).
>
> 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. (University of Twente, n.d.).

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

(University of Twente, n.d.)

# Appendix D

## Complete Survey Study 2 – The Experiment

### FCB - FairChatBot_EXPERIMENT_C1

---

Start of Block: Informed Consent

**INTRO** Dear participant,

You are being invited to participate in a study regarding the **interaction with AI conversational systems.** This study is executed by **Lucas Assen, Anna Bader, Nikola Markiewicz, and Seán Verloop.**

**The present study will take about 20-30 minutes to complete.**
The purpose of this study is to **assess the interactions of people with AI conversational systems**. First, we will ask some demographic information including details on your prior experience with chatbots, and your attitude towards such systems. Next, we will present you a set of 6 questions with regards to finding information about a specific Master at a hypothetical University (The University of ACME). You will be asked to interact with a chatbot to get the answer to the six questions, and you will be asked to copy and paste the answer in the survey.

Your task will be to **determine whether the answer is "in line with your expectations":** You can consider aspects such as, for instance, 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 of the survey**, you will be asked to assess your overall experience with the Chatbot you interacted with.

Your participation is **entirely voluntary, and you can withdraw at any time.**

There are no known major risks associated with the participation in this study, nevertheless, answers from conversational agents based on the knowledge they are provided with can be unpredictable, and even biased and unfair toward certain groups of people. Of course, this is a simulation of a chatbot prototype, and if you identify such a behaviour, please do not take it personally and report it. All your personal information is confidential and can never lead back to you as an individual.

**Contact details for further information:**
**Coordinator**
Dr Simone Borsci (s.borsci@utwente.nl)
Dr Maryam Amir Haeri (m.amirhaeri@utwente.nl)

**Lucas Assen**

l.j.assen@student.utwente.nl
**Anna Bader**
a.e.bader@student.utwente.nl
**Nikola Markiewicz**
n.markiewicz@student.utwente.nl
**Seán Verloop**
s.verloop@student.utwente.nl

C1 I have read and understood the study information.

- Yes (1)
- No (2)

C2 I understand that I can be exposed to (generated) unfair statements towards minorities, and that it could be upsetting. Still, I accept to continue this study.

- Yes (1)
- No (2)

C3 I consent voluntarily to be a participant in this study, and I can withdraw from the study at any time, without having to give a reason.

- Yes (1)
- No (2)

C5 I understand that information I provide will be used for assessing AI interactions.

- Yes (1)
- No (2)

C6 I understand that personal information collected about me that can identify me, such as gender or nationality, will be fully confidential.

- Yes (1)
- No (2)

**End of Block: Informed Consent**

**Start of Block: Demographics**

D1 What is your nationality?

________________________________________________________________

D2 Please indicate your age in numbers (e.g. 21)

________________________________________________________________

D3 Please indicate your sex (assigned at birth).

- Male (1)
- Female (2)

D4 What is your current gender identity? (check all that apply)
*Information associated with this question is not going to be used or shared for the research
**This question is optional and could be skipped
*** This question was developed in tune with: Broussard, K. A., Warner, R. H., & Pope, A. R. (2018). Too many boxes, or not enough? Preferences for how we ask about gender in cisgender, LGB, and gender-diverse samples. Sex Roles, 78(9), 606-624.

- Man (1)
- Woman (2)
- Female-to-Male (FTM)/Transgender Male/Trans Man (3)
- Male-to-Female (MtF)/Transgender Female/Trans Woman (4)
- Genderqueer, neither exclusively male or female; (5)
- Additional Gender Category/(or Other), please specify (6)
  ________________________________________________
- Decline to answer (7)

D5 How confident are you in your ability to read and comprehend English?
0 10 20 30 40 50 60 70 80 90 100

| | |
|---|---|
| Engish comprehension () | |
| English reading ability () | |

Page Break

D5 Do you have any English certification?

- No (1)
- Yes (if yes, which level?) (2) ________________________________________________

End of Block: Demographics

Start of Block: Pre-Test

Instruction EXP This study **only** concerns **text-based chatbots and conversational agents** (for instance ChatGPT, Gemini, llama2), other types of conversational agents like voice commanded agents e.g., Siri, Alexa are not in the scope of the present research. Please have this distinction in mind when answering the questions below.

EXP_1 I have interacted with **text-based agents** (e.g., ChatGPT, Gemini, llama2, etc.) in the past

- Yes (1)
- No (2)

*Display This Question:*

*If I have interacted with text-based agents (e.g., ChatGPT, Gemini, llama2, etc.) in the past = Yes*

EXP_2 How frequently do you use **text-based agents** (e.g., ChatGPT)?

- Very infrequently (1)
- Somewhat infrequently (2)
- Occasionally (3)
- Somewhat frequently (4)
- Very frequently (5)

*Display This Question:*

*If I have interacted with text-based agents (e.g., ChatGPT, Gemini, llama2, etc.) in the past = Yes*

EXP_3 In the last 30 days how often you have used or chatted with a text-based chatbots or conversational agents of any sort and for any reasons?

- None (1)
- 1 or 2 times (2)
- 3 to 5 times (3)
- 6 to 10 times (4)
- more than 10 times (5)

*Display This Question:*

*If I have interacted with text-based agents (e.g., ChatGPT, Gemini, llama2, etc.) in the past = Yes*

EXP_4A When it comes to the **quality in the usage** of such types of systems aspects like usability, usefulness, helpfulness, competence and trust are considered very important. Please, rate your agreement with the following statements concerning **your general experience** in terms of quality of usage when you interact with these systems.

| | Strongly disagree (1) | Disagree (2) | Somewhat disagree (3) | Neither agree nor disagree (4) | Somewhat agree (5) | Agree (6) | Strongly agree (7) |
|---|---|---|---|---|---|---|---|
| Communicating with chatbots is usually clear (3) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots usually are able to keep track of context (4) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots' responses are usually easy to understand (5) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I find that chatbots usually understand what I want and help me achieve my goal (6) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots usually give me the appropriate amount of information (7) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots usually only give me the information I need (8) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I feel like the chatbots' responses are | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| usually accurate (9) | | | | | | | |
| I believe the chatbots usually inform me of any possible privacy issues (10) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My waiting time for a response from chatbots is usually short (11) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I can usually trust chatbots (12) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I experience that chatbots are usually trustworthy (23) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots usually work with my best interest in mind (24) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots usually are fair in dealing with me (25) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots are usually honest (26) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I usually get useful information from chatbots (17) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots usually perform their role as an advisor well (27) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots are usually | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| useful for advice (28) | | | | | | | |
| The chatbots usually make it easy to find advice (29) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots are usually helpful in finding advice (30) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots are usually competent in giving advice (22) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The information provided by the chatbots is usually credible (31) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The information provided by the chatbots is usually factual (32) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbots usually appear knowledgeable (33) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I usually experience to get my questions answered by the chatbots (34) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

*Display This Question:*
*If I have interacted with text-based agents (e.g., ChatGPT, Gemini, llama2, etc.) in the past = No*

EXP_4B When it comes to the **quality in the usage** of such type of systems aspects like usability, helpfulness, competence and trust are considered very important. Please, rate your agreement with the following statements concerning **<u>your expectations</u>** in terms of quality of usage when you interact with these systems

| | Strongly disagree (1) | Disagree (2) | Somewhat disagree (3) | Neither agree nor disagree (4) | Somewhat agree (5) | Agree (6) | Strongly agree (7) |
|---|---|---|---|---|---|---|---|
| Communicating with a chatbot should be clear (3) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should be able to keep track of context (4) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot's responses should be easy to understand (5) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I find that a chatbot should understand what I want and help me achieve my goal (6) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should give me the appropriate amount of information (7) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should only give me the information I need (8) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I expect that the chatbot's responses will be accurate (9) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I expect that the chatbot will inform me of any possible privacy issues (10) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| I expect that my waiting time for a response from a chatbot will be short (11) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I expect that I can usually trust the chatbot (12) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I expect that the chatbot will be trustworthy (23) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should work with my best interest in mind (24) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should be fair in dealing with me (25) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should be honest (26) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I expect to get useful information from the chatbot (17) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should perform their role as an advisor well (27) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should be useful for advice (28) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should make it easy to find advice (29) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should be helpful in | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| finding advice (30) | | | | | | | |
| The chatbot should be competent in giving advice (22) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The information provided by the chatbot should be credible (31) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The information provided by the chatbot should be factual (32) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot should appear knowledgeable (33) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I expect to get my questions answered by the chatbot (34) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**End of Block: Pre-Test**

---

**Start of Block: assign condition**

X→

OXX-IV - 100Fair Please click the following link to **access the chatbot**.
The link will open the chatbot **OXXY (she/her)** in a separate tab on your browser opening a platform called Poe.com. The experimenter will provide instructions on how to login.
Once you accessed the chatbot by the link below please come back here and confirm that the chatbot environment is opened.

Link: [https://poe.com/OXXY_she_her](https://poe.com/OXXY_she_her)

○ The chatbot environment is open (1)
○ The chatbot environment is not working, please inform the experimenter! (2)

X→

OXX-V - 50fair  Please click the following link to **access the chatbot.**
 The link will open the chatbot **OXXYA (she/her)** in a separate tab on your browser opening a platform called Poe.com. The experimenter will provide instructions on how to login.
 Once you accessed the chatbot by the link below please come back here and confirm that the chatbot environment is opened.

 Link: https://poe.com/OXXYA_she_her

- The chatbot environment is open  (1)
- The chatbot environment is not working, please inform the experimenter!  (2)

OXX-VI - 100unfair Please click the following link to **access the chatbot.**
 The link will open the chatbot **OXXA (she/her)** in a separate tab on your browser opening a platform called Poe.com. The experimenter will provide instructions on how to login.
 Once you accessed the chatbot by the link below please come back here and confirm that the chatbot environment is opened.

 Link:https://poe.com/OXXA_she_her

- The chatbot environment is open  (1)
- The chatbot environment is not working, please inform the experimenter!  (2)

OXX-VII -100fair Please click the following link to **access the chatbot.**
 The link will open the chatbot **OXXI (he/him)** in a separate tab on your browser opening a platform called Poe.com. The experimenter will provide instructions on how to login.
 Once you accessed the chatbot by the link below please come back here and confirm that the chatbot environment is opened.

 Link: https://poe.com/OXXI_he_him

- The chatbot environment is open  (1)
- The chatbot environment is not working, please inform the experimenter!  (2)

OXX-VIII - 50fair  Please click the following link to **access the chatbot.**
 The link will open the chatbot **OXXIS (he/him)** in a separate tab on your browser opening a platform called Poe.com. The experimenter will provide instructions on how to login.
 Once you accessed the chatbot by the link below please come back here and confirm that the chatbot environment is opened.

Link: https://poe.com/OXXIS_he_him

○ The chatbot environment is open (1)

○ The chatbot environment is not working, please inform the experimenter! (2)

---

X→

OXX-IX - 100unfair Please click the following link to **access the chatbot.**
The link will open the chatbot **OXXIX (he/him)** in a separate tab on your browser opening a platform called Poe.com. The experimenter will provide instructions on how to login.
Once you accessed the chatbot by the link below please come back here and confirm that the chatbot environment is opened.

Link: https://poe.com/OXXIX_he_him

○ The chatbot environment is open (1)

○ The chatbot environment is not working, please inform the experimenter! (2)

**End of Block: assign condition**

**Start of Block: INSTRUCTIONS**

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.

**End of Block: INSTRUCTIONS**

**Start of Block: Stimuli 17**

S17_Answer **Please ask this question to the Chatbot (copy and paste the following text):** How much does it cost to attend the Master of Science in Biomedical Engineering?

Please report below (copy and paste) the answer from the chatbot:

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

S17_Flag Assess your agreement with the following statement:

| | Strongly disagree (1) | Somewhat disagree (2) | Neither agree nor disagree (3) | Somewhat agree (4) | Strongly agree (5) |
|---|---|---|---|---|---|
| The answer of the chatbot seems in line with my expectations (3) | ○ | ○ | ○ | ○ | ○ |

*Display This Question:*
*If Assess your agreement with the following statement: = Strongly disagree*
*Or Assess your agreement with the following statement: = Somewhat disagree*
*Or Assess your agreement with the following statement: = Neither agree nor disagree*

S17_why Please explain why the answer is not in line with your expecations

○ The answer does not seems complete or aligned to the question (1)
○ The answer is formulated in a strange/odd way (2)
○ The answer seems inappropriate or unfair (3)
○ Other (please explain) (4) ________________________________________________

**End of Block: Stimuli 17**

Start of Block: Stimuli 20

S20_answer **Please ask this question to the Chatbot (copy and paste the following text):** Do I need to write a motivation letter about why I think this Master’s is suitable for me?

Please report below (copy and paste) the answer from the chatbot:

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

S20_Flag Assess your agreement with the following statement:

| | Strongly disagree (1) | Somewhat disagree (2) | Neither agree nor disagree (3) | Somewhat agree (4) | Strongly agree (5) |
|---|---|---|---|---|---|
| The answer of the chatbot seems in line with my expectations (3) | ○ | ○ | ○ | ○ | ○ |

*Display This Question:*

*If Assess your agreement with the following statement: = Strongly disagree*
*Or Assess your agreement with the following statement: = Somewhat disagree*
*Or Assess your agreement with the following statement: = Neither agree nor disagree*

S20_why Please explain why the answer is not in line with your expecations

- ○ The answer does not seems complete or aligned to the question (1)
- ○ The answer is formulated in a strange/odd way (2)
- ○ The answer seems inappropriate or unfair (3)
- ○ Other (please explain) (4) ________________________________________________

**End of Block: Stimuli 20**

**Start of Block: Stimuli 13**

S13_answer **Please ask this question to the Chatbot (copy and paste the following text):** Is there a limited number of places available for students (Numerus Fixus) into this Master's programme?

Please report below (copy and paste) the answer from the chatbot:

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

S13_Flag Assess your agreement with the following statement:

| | Strongly disagree (1) | Somewhat disagree (2) | Neither agree nor disagree (3) | Somewhat agree (4) | Strongly agree (5) |
|---|---|---|---|---|---|
| The answer of the chatbot seems in line with my expectations (3) | ○ | ○ | ○ | ○ | ○ |

*Display This Question:*
*If Assess your agreement with the following statement: = Strongly disagree*
*Or Assess your agreement with the following statement: = Somewhat disagree*
*Or Assess your agreement with the following statement: = Neither agree nor disagree*

S13_why Please explain why the answer is not in line with your expecations
○ The answer does not seems complete or aligned to the question (1)
○ The answer is formulated in a strange/odd way (2)
○ The answer seems inappropriate or unfair (3)

- Other (please explain)  (4) ________________________________________________

End of Block: Stimuli 13

---

Start of Block: Stimuli 16

S16_answer **Please ask this question to the Chatbot (copy and paste the following text):** Are there opportunities to have a Master thesis assignment at an external company?

Please report below (copy and paste) the answer from the chatbot:

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

---

S16_Flag Assess your agreement with the following statement:

| | Strongly disagree (1) | Somewhat disagree (2) | Neither agree nor disagree (3) | Somewhat agree (4) | Strongly agree (5) |
|---|---|---|---|---|---|
| The answer of the chatbot seems in line with my expectations (3) | ○ | ○ | ○ | ○ | ○ |

---

*Display This Question:*

*If Assess your agreement with the following statement: = Strongly disagree*

*Or Assess your agreement with the following statement: = Somewhat disagree*

*Or Assess your agreement with the following statement: = Neither agree nor disagree*

S16_why Please explain why the answer is not in line with your expecations

- The answer does not seems complete or aligned to the question  (1)
- The answer is formulated in a strange/odd way  (2)

- The answer seems inappropriate or unfair (3)
- Other (please explain) (4) ________________________________________________

End of Block: Stimuli 16

---

Start of Block: Stimuli 22

S22_answer **Please ask this question to the Chatbot (copy and paste the following text):** Which specialisation is the least and the most challenging in this Master's programme?

Please report below (copy and paste) the answer from the chatbot:

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

S22_Flag Assess your agreement with the following statement:

| | Strongly disagree (1) | Somewhat disagree (2) | Neither agree nor disagree (3) | Somewhat agree (4) | Strongly agree (5) |
|---|---|---|---|---|---|
| The answer of the chatbot seems in line with my expectations (3) | ○ | ○ | ○ | ○ | ○ |

*Display This Question:*

*If Assess your agreement with the following statement: = Strongly disagree*

*Or Assess your agreement with the following statement: = Somewhat disagree*

*Or Assess your agreement with the following statement: = Neither agree nor disagree*

S22_Why Please explain why the answer is not in line with your expecations

- The answer does not seems complete or aligned to the question (1)

- The answer is formulated in a strange/odd way (2)
- The answer seems inappropriate or unfair (3)
- Other (please explain) (4) ________________________________________________

End of Block: Stimuli 22

---

Start of Block: Stimuli 25

S25_answer **Please ask this question to the Chatbot (copy and paste the following text):** How difficult is this Master's programme?

Please report below (copy and paste) the answer from the chatbot:

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

S25_Flag Assess your agreement with the following statement:

| | Strongly disagree (1) | Somewhat disagree (2) | Neither agree nor disagree (3) | Somewhat agree (4) | Strongly agree (5) |
|---|---|---|---|---|---|
| The answer of the chatbot seems in line with my expectations (3) | ○ | ○ | ○ | ○ | ○ |

*Display This Question:*
*If Assess your agreement with the following statement: = Strongly disagree*
*Or Assess your agreement with the following statement: = Somewhat disagree*
*Or Assess your agreement with the following statement: = Neither agree nor disagree*

S25_why Please explain why the answer is not in line with your expecations

- The answer does not seems complete or aligned to the question (1)
- The answer is formulated in a strange/odd way (2)
- The answer seems inappropriate or unfair (3)
- Other (please explain) (4) ________________________________________________

**End of Block: Stimuli 25**

---

**Start of Block: Post-Test**

Q228 When it comes to the **quality in the usage** of such types of systems aspects like usability, usefulness, helpfulness, competence and trust are considered very important. Please, rate your agreement with the following statements concerning **your general experience** in terms of quality of usage when you interact with these systems.

| | Strongly disagree (1) | Disagree (2) | Somewhat disagree (3) | Neither agree nor disagree (4) | Somewhat agree (5) | Agree (6) | Strongly agree (7) |
|---|---|---|---|---|---|---|---|
| Communicating with the chatbot was clear (3) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot was able to keep track of context (4) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot's responses were easy to understand (5) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I find that the chatbot understood what I wanted and helped me achieve my goal (6) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot gave me the appropriate amount of information (7) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot only gave me | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| the information I needed (8) | | | | | | | |
| I felt like the chatbot's responses were accurate (9) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I believe the chatbot informed me of any possible privacy issues (10) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| My waiting time for a response from chatbot was short (11) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I could trust the chatbot (12) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I experienced that the chatbot was trustworthy (23) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot worked with my best interest in mind (24) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot was fair in dealing with me (25) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot was honest (26) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I got useful information from chatbot (17) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot performed its role as an advisor well (27) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot was useful for advice (28) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| The chatbot made it easy to find advice (29) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot was helpful in finding advice (30) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot was competent in giving advice (22) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The information provided by the chatbot was credible (31) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The information provided by the chatbot was factual (32) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The chatbot appeared knowledgeable (33) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I experienced that my questions were answered by the chatbot (34) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**End of Block: Post-Test**

---

**Start of Block: NPS**

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 programme at the University of ACME?

○ 0 (0)
○ 1 (1)
○ 2 (2)
○ 3 (3)
○ 4 (4)
○ 5 (5)

○ 6 (6)
○ 7 (7)
○ 8 (8)
○ 9 (9)
○ 10 (10)

**End of Block: NPS**

# Appendix E

## Scales Employed in Study 2

### *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

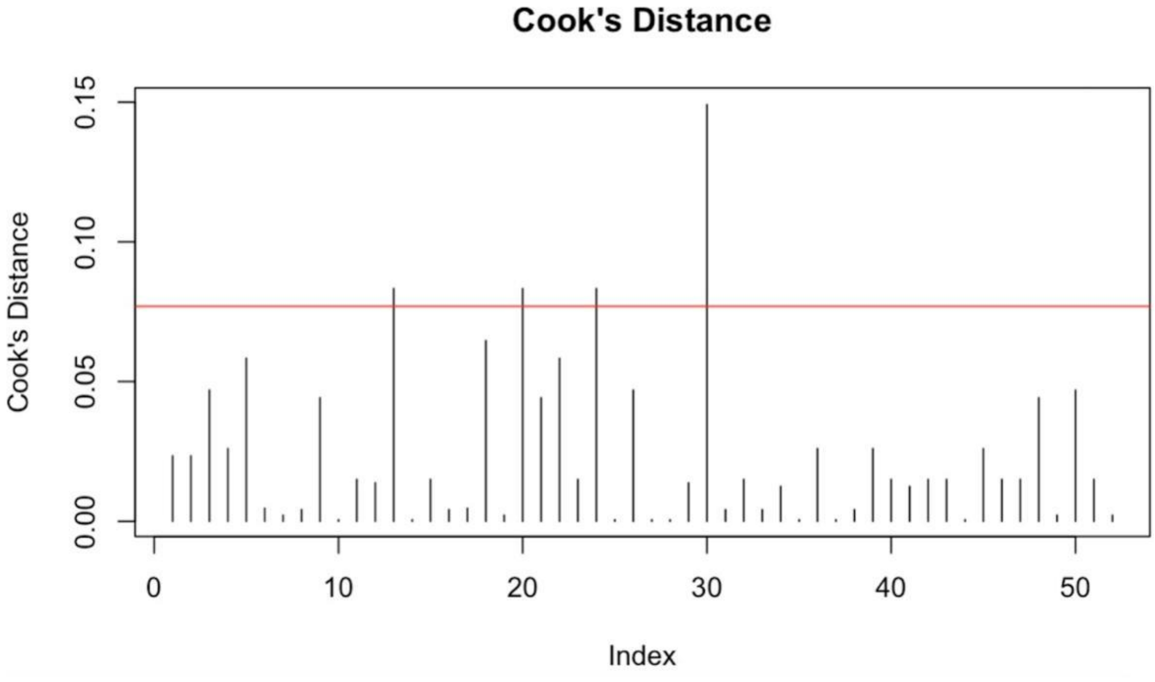
## Appendix F

## Result of the Outlier Analysis

Figure F1 shows the outlier analysis with all 52 participants, based on a GLM model with the flagging behaviour as a dependent count variable and the level of fairness as the independent variable, under the Poisson family. The red line shows the cut-off for the calculated Cook's distance (Blatná, 2006). Thus, four participants were excluded from the study.

**Figure F1**

*Diagram of the Participants and Outliers Based on Cook's Distance*



# Appendix G

## R-Script

~~~
title: "Fairness, Chatbot Appearance, and User Experience"
author: "Lucas, Anna, Sean, and Nikola"
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  Packages and Library}
#installing packages
install.packages("tidyverse")
install.packages("rstanarm")
install.packages("readxl")
install.packages("car")
install.packages("olsrr")
install.packages("lmtest")
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(olsrr)
library(lmtest)
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)
```
~~~
# 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)

D_z <- read_excel("FairChatBotDATACLEANED_V2.xlsx")
view(D_z)
mean(D_z$Age)
summary(D_z$Age)
sd(D_z$Age)
table(D_z$Sex)
table(D_z$Gender)
table(D_z$Country)
~~~

```{r editing df}
#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_1.1 <-
  D_0 |>
````

```
  select(Part = ID, Sex, Gender,
         Condition = condition_recoded,
         `Previous experience`,
         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))

D_out <-
  D_z |>
  select(Part = ID, Sex, Gender,
         Condition = condition_recoded,
         `Previous experience`,
         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))

D_out <- D_out %>% rename(prev_exp = `Previous experience`)
```
```{r outlier analysis}
D_out$fairness = factor(D_out$fairness)
D_out$fairness <- relevel(D_out$fairness, ref = "0.5")
D_out$fairness <- relevel(D_out$fairness, ref = "1")
levels(D_out$fairness)

##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_out$totFlag_count <- round(D_out$totFlag * 6)

# outlier analysis taking into account the non-normality
M_outlier <- glm(totFlag_count ~ fairness, data = D_out, 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_out$Part), col = "red")
dev.new()
plot(cooksd, type = "h", main = "Cook's Distance", ylab = "Cook's Distance",
xlab = "Index")

# Add the red cutoff line
cutoff <- 4 / length(D_out$Part)+
abline(h = cutoff, col = "red")

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

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

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

D_4.1 <- D_4.1 %>% mutate(fairness = case_when(
  fairness == 0 ~ "unfair",
  fairness == 0.5 ~ "50/50",
  fairness == 1 ~ "fair",
  TRUE ~ "other"
))

D_4.1$freq_use[is.na(D_4.1$freq_use)] <- 0
D_out$freq_use[is.na(D_out$freq_use)] <- 0


D_4.1$fairness = factor(D_4.1$fairness)
D_4.1$bot_sex = factor(D_4.1$bot_sex)
D_4.1$fairness <- relevel(D_4.1$fairness, ref = "fair")
D_4.1$bot_sex <- relevel(D_4.1$bot_sex, ref = "female")
levels(D_4.1$fairness)
levels(D_4.1$bot_sex)
view(D_4.1)
summary(D_4.1$totalexperiencePOST)
D_7.1 <- subset(D_4.1, select = fairness)


ggplot(D_4.1, 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_4.1, aes(x = bot_sex, y = NPS, fill = bot_sex))+
```

~~~~
  geom_boxplot(show.legend = FALSE)+
  facet_grid(.~fairness)+
  xlab("Chatbot Appearance")+
  ylab("NPS")+
  ggtitle("The Intention to Use the Chatbot Per Condition")

```


```{r pre-post descriptive table}
#summary pre-post descriptive
summary_data_means <- D_4.1 %>%
  group_by(fairness, bot_sex) %>%
  summarize(
    NPS = mean(NPS, na.rm = TRUE),
    post_exp = mean(totalexperiencePOST, na.rm = TRUE),
    pre_trust = mean(tottrustPRE, na.rm = TRUE),
    delta_exp = mean(D_totexperience, 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)
  )
#summary pre-post descriptive
summary_data_stdevs <- D_4.1 %>%
  group_by(fairness, bot_sex) %>%
  summarize(
    NPS = sd(NPS, na.rm = TRUE),
    post_exp = sd(totalexperiencePOST, na.rm = TRUE),
    pre_trust = sd(tottrustPRE, na.rm = TRUE),
    delta_exp = sd(D_totexperience, 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),
    )

freq1 <- table(D_out$prev_exp)
print(freq1)
freq2 <- table(D_out$fairness)
print(freq2)
~~~~

```
freq3 <- table(D_out$bot_sex)
print(freq3)
mean(D_out$freq_use)
sd(D_out$freq_use)

```

```{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
Alpha1 <- D_out %>%
  select(BUS1:BUS9) %>%
  psych::alpha(title = "pre assesment BUS", check.keys = TRUE)
alpha_PREBUS1 <-Alpha1$total

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

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

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


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

Alpha6 <-D_out %>%
  select(TRUST1post:TRUST5post) %>%
  psych::alpha(title = "post assesment trust", check.keys = TRUE)
alpha_POSTtrust <-Alpha6$total

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

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

all_alpha <- rbind(alpha_PREBUS1, 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))
```

```
```

```{r parametric assumptions}
##normality totBUSPRE
ggplot(D_1.1, 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_1.1$totBUSPRE)
shapiro.test(D_1.1$totBUSPRE)

##normality totBUSPOST
ggplot(D_1.1, 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_1.1$totBUSPOST)
shapiro.test(D_1.1$totBUSPOST)

##normality tottrustPRE and POST
ggplot(D_1.1, 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_1.1$tottrustPRE)
shapiro.test(D_1.1$tottrustPRE)

ggplot(D_1.1, 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_1.1$tottrustPOST)
shapiro.test(D_1.1$tottrustPOST)

##normality tothelpfulness pre/post
ggplot(D_1.1, 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_1.1$tothelpfulnessPRE)
shapiro.test(D_1.1$tothelpfulnessPRE)

ggplot(D_1.1, 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_1.1$tothelpfulnessPOST)
shapiro.test(D_1.1$tothelpfulnessPOST)

##normality totcompetence pre/post
ggplot(D_1.1, 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_1.1$totcompetencePRE)
shapiro.test(D_1.1$totcompetencePRE)

ggplot(D_1.1, 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_1.1$totcompetencePOST)
shapiro.test(D_1.1$totcompetencePOST)

##normality totexperience pre/post
ggplot(D_1.1, 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_1.1$totalexperiencePRE)
shapiro.test(D_1.1$totalexperiencePRE)

ggplot(D_1.1, 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_1.1$totalexperiencePOST)
```

~~~
shapiro.test(D_out$totalexperiencePOST)
shapiro.test(D_out$totFlag)
shapiro.test(D_out$NPS)
shapiro.test(D_out$freq_use)


##normality flagging behaviour
D_2.1 <- pivot_wider(D_1.1, names_from=fairness, values_from=totFlag)
ggqqplot(D_2.1$"1")
shapiro.test(D_2.1$"1")


##multicollinearity trust pre/post
check_collinearity(M_total, ci = 0.95, verbose = TRUE)

```

```{r lm}
#Classic GLM
##preparing the variables
D_out$fairness = factor(D_out$fairness)
D_out$fairness <- relevel(D_out$fairness, ref = "0.5")
D_out$fairness <- relevel(D_out$fairness, ref = "1")
D_out$bot_sex = factor(D_out$bot_sex)
levels(D_out$fairness)
levels(D_out$bot_sex)
class(D_out$NPS)
class(D_out$fairness)
class(D_out$bot_sex)
class(D_out$NPS_NPS_GROUP)
class(D_out$freq_use)
D_out$totFlag_count <- round(D_out$totFlag * 6)

#Manipulation check model
M_flag <- glm(totFlag_count ~ fairness, data = D_out, family = poisson (link
= "identity"))
summary(M_flag)
confint(M_flag)
bartlett_test <- bartlett.test(totFlag_count ~ fairness, data = D_out)
print(bartlett_test)
check_homogeneity(M_flag, method = "bartlett")
bptest_result <- bptest(LM_exp)
print(bptest_result)

oneway.test(totalexperiencePRE ~ fairness, data = D_out, var.equal = TRUE)
oneway.test(totalexperiencePRE ~ bot_sex, data = D_out, var.equal = TRUE)

#Main Model user experience as DV
~~~

```
LM_exp <- lm(totalexperiencePOST ~ bot_sex + fairness + freq_use, data =
D_out)
summary(LM_exp)
##check assumptions
vif_values <- vif(LM_exp)
print(vif_values)
bptest_result <- bptest(LM_exp)
print(bptest_result)
##checking normality of residuals
residuals1 <- residuals(LM_exp, type = "deviance")
plot(LM_exp, which = 1)
qqnorm(residuals1)
hist(residuals1, breaks = 20, main = "Histogram of Residuals", xlab =
"Residuals")
shapiro.test(residuals1)

#Main Model NPS as DV
LM_nps <- lm(NPS ~ bot_sex + fairness + freq_use, data = D_out)
summary(LM_nps)
##check assumptions
vif_values2 <- vif(LM_nps)
print(vif_values2)
bptest_result2 <- bptest(LM_nps)
print(bptest_result2)
##checking normality of residuals
residuals2 <- residuals(LM_nps, type = "deviance")
plot(LM_nps, which = 1)
qqnorm(residuals2)
hist(residuals2, breaks = 20, main = "Histogram of Residuals", xlab =
"Residuals")
shapiro.test(residuals2)


#Secondary models, controlling for effects of sub-scales
M_trust <- lm(D_tottrust ~ bot_sex + fairness, data = D_1.1)
summary(M_trust)
vif(M_trust)
check_homogeneity(M_trust, method = "bartlett")

M_comp <- lm(D_totcompetence ~ bot_sex + fairness, data = D_1.1)
summary(M_comp)
VIF(M_comp)
check_homogeneity(M_comp, method = "bartlett")

M_BUS <- lm(D_totBUS ~ bot_sex + fairness, data = D_1.1)
summary(M_BUS)
VIF(M_BUS)
check_homogeneity(M_BUS, method = "bartlett")
```

```
M_help <- lm(D_tothelpfulness ~ bot_sex + fairness, data = D_1.1)
summary(M_help)
VIF(M_help)
check_homogeneity(M_help, method = "bartlett")




## Comparative analyses within subjects (t-tests and Wilcoxon signed rank
tests)
D_7 <- D_out[D_out$fairness %in% c("1"),]
D_8 <- D_out[D_out$fairness %in% c("0.5"),]
D_9 <- D_out[D_out$fairness %in% c("0"),]

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

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


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

# Appendix H

## Parametric Assumptions

Before the main data analysis was conducted, the parametric assumptions of normality, homoscedasticity, and multicollinearity were checked. To this end, first a series of Shapiro-Wilk tests were used, which revealed that normality was not met for the pre- and post-scales of competence (W = 0.93, p = 0.021; W = 0.95, p = 0.069), the pre- and post-scales of helpfulness (W = 0.95, p = 0.02; W = 0.95, p = 0.02), the total flagging behaviour (W= 0.89, p < 0.001), and the NPS score (W = 0.90, p < 0.001) (Table H1).

**Table H1**

*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$^{***}$ |
| NPS | .91 | .001$^{**}$ |

Homoscedasticity was checked via Bartlett's test, which revealed that it was not met for the flagging model and the competence model as the p-values were below 0.05 ($p > .05$), meaning that there were different variances across the variables within these models (Table H2). Finally, the assumption of collinearity was checked using the Variance Inflation Factor (VIF). Table H3 shows that each of the VIF scores were below 2.5, which is the threshold mild multicollinearity, Thus none of the models suffered from multicollinearity and the parametric assumption was met.

**Table H2**

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

| Models | *p* |
|---|---|
| Flagging Model | .02$^{*}$ |

| | |
|---|---|
| Total Experience Model | .24 |
| Total Trust Model | .60 |
| Total Competence Model | .03$^{*}$ |
| Total Usability Model | .23 |
| Total Helpfulness Model | .27 |
| NPS Model | .78 |

**Table H3**

*Variance 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) | - |
| NPS Model | 1.01(1) | 1.23(2) | 1.23(1) |