Perceivable Unfairness of Conversational Agents: Implications for Trust, Competence, Helpfulness and Usability

Nikola Markiewicz (s2716593)

Department of Behavioural, Management and Social Sciences, University of Twente

202000383: Bachelor Thesis Human Factors and Engineering Psychology

Dr. Simone Borsci

June 25, 2024

Word count: 9997

Abstract

In the contemporary society, chatbots, a type of AI-based conversational agents (CAs) are one of the vastly developing technologies utilised in multiple domains such as marketing, academia, or customer service. As with any system, there are potential risks associated with using chatbots that can spread inaccurate or biassed (unfair) information. The spreading of bias and misinformation puts minority groups at risk of facing discrimination due to their individual characteristics. On top of that, previous research suggests that male and female chatbots tend to be treated and rated differently. Thus, this study focused on the effect of the level of unfairness and appearance of the chatbot (sex: female vs male) on the quality of the interaction measured through perceived trust, helpfulness, competence, and usability. To test this, first, a pilot study was conducted to create and validate explicit negatively biassed sentences that were used in the main phase of the experiment as chatbot knowledge. Participants interacted with a chatbot and reported their attitude and experience ratings before and after the interaction by filling items about trust, usability, helpfulness and competence. A pre-post, between-subjects design 2 (appearance of the chatbot: male or female) by 3 (level of hallucinations: 100% fair, 50% fair/unfair, or 100% unfair) was employed. The main findings show a significant effect of unfairness (p <.001) and no effect of appearance (p = .267). Significant differences between the pre and post-measurements were identified in the fair and unfair conditions. These results might suggest that people anticipate and thus tolerate a certain level of unfairness in chatbots. When the chatbot is fair, expectations are exceeded, and post ratings are significantly different. When the chatbot is unfair, people tolerate it to a certain extent, as there is no significant difference between pre and post-measures of usability and helpfulness. There is, however, a significant decrease in perceived trust and competence. Future research should further validate the stimuli created during the pilot, manipulate the expressions of chatbot appearance, the effect of previous experience and expectations, and possibly manipulate the levels of fairness by adding more conditions.

*Keywords***:** chatbot, fairness, appearance, artificial intelligence, AI hallucinations, usability, fairness, competence, helpfulness

Introduction

Conversational agents (CAs) are dialogue systems capable of receiving written or spoken input, processing it, and generating a natural-language response for its users (Allouch et al., 2021). The more advanced systems operate based on a large language model (LLM), which is a generative mathematical algorithm capable of using and tokenizing individual characters and words from large data sets to recognize, summarise, translate, predict, and generate content (Shanahan, 2024). In principle, they generate the most statistically likely sequence of words in response to the user's inquiries. LLMs are a form of generative artificial intelligence (AI) capable of creating new original content such as text, imagery, audio, or synthetic data (Zhao et al., 2024).

CAs that interact with their users through text are called chatbots (Allouch et al., 2021). Chatbots are utilised in multiple domains such as academia (Zhao et al., 2024), mental health care (Balan et al., 2024) or customer service (Chakrabarti & Luger, 2015) and shape society by changing how people search for information, study, work, and interact with each other. The level of autonomy a chatbot has can vary from non-autonomous ones that fully depend on the decisions of humans behind them, semi-autonomous ones that have certain freedoms but still require human management, up to fully autonomous ones that can operate on their own (Balan et al., 2024).

We live in the era of the "chatbot tsunami" (Grudin & Jacques, 2019) as their use and popularity have increased gradually over the past years. As reported by Vyshnevska (2024), 58% of B2B companies and 42% of B2C companies make use of chatbots on their websites. Kaminska (2023) predicts the chatbot market worldwide will reach around 1.25 billion U.S. dollars, which is a significant increase from 2016 when it was valued at 190.8 million U.S.

dollars. These numbers alone reflect the gradual increase in the popularity of conversational agents.

Undoubtedly, there are many advantages to using conversational agents, such as easy access and general ease of use, short response time and decreased pressure on human workers (Even et al., 2022). However, it is crucial to consider the potential negative consequences and threats of using these tools, such as the correctness and reliability of the outputs they provide.

Dialogue systems such as conversational agents can exhibit a phenomenon called AI hallucination “wherein a large language model (LLM) - often a generative AI chatbot or computer vision tool - perceives patterns or objects that are non-existent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate” (*What Are AI Hallucinations? | IBM*, n.d.). It is estimated that depending on the specific model, chatbots hallucinate at least 3% of their answers up to even 27% (Hughes, 2024). Hallucinations can vary from potentially believable information to absolute gibberish. For instance, if the user asks a chatbot when Leonardo Da Vinci painted the Mona Lisa, they might receive an incorrect AI-generated response stating the painting was created in 1815. Meanwhile, we know that the Mona Lisa was most probably created between 1503 and 1506, perhaps continuing even until 1517.

Apart from providing factually incorrect answers, hallucinations can generate unfair or biassed information and contribute to spreading harmful stereotypes that target various minorities, such as gender bias. Caliskan and colleagues (2017) report that semantics that conversational agents derive from large language data sets contain human-like biases. The text corpora available to the agents include various types of biases, including historical, gender, or racial biases among others. Furthermore, chatbots can exhibit algorithmic biases generated by the model itself during a biassed decision-making process due to the system’s training (Wang et al., 2023). Notably, LLM utilises large data sets that include outdated, fabricated or biassed information that is not verified. Furthermore, many errors produced by LLM are almost imperceptible due to the high plausibility of the provided information, which makes it a challenge for both systems and humans to identify the hallucinations (Zhang et al., 2023).

Current research on conversational agents has covered multiple topics such as acceptance and use of the technology, experience in terms of satisfaction, trust and engagement, the emotional impact, potential downsides such as distribution of false information, and humaneness of chatbots (Rapp et al., 2021). However, so far no one has studied how the unfairness of chatbots may influence these attitudes and experiences of users. They have operated under the assumption that chatbots are fair even though they may provide incorrect information. There is no universal definition of unfairness, as it is highly context dependent. Hardmeier et al. (2021) propose that a "successful biassed statement is one that clarifies the harm it causes and who suffers from it, and includes statements that paint a social group unfavourably, degrade it, or deny its existence" (p. 2) In this paper, we define unfairness as the lack of equality and justice. Thus, an unfair output from a chatbot is one that puts either a person or a minority group in an unequal position in comparison to other individuals based on characteristics such as gender, skin tone, nationality, or others. Contrarily, a fair output represents all individuals and minority groups in an equal manner.

In the field of human-computer interaction, we recognize the Computers Are Social Actors (CASA) paradigm which facilitates understanding of how people perceive and interact with technologies. According to this perspective, we tend to apply the same social heuristics as we do in social interactions to computers (Nass et al., 1994). This has interesting implications for the relationship between users and conversational agents. These CAs are being perceived as social actors that collaborate with us in the process of searching for information or making decisions. Thus, similarly to interaction with humans, if we are being lied to, like by receiving false or unfair information or omitting some facts, we lose the trust we initially had for the person which significantly affects our relationship. Similarly, if CAs provide false or unfair information, the entire collaboration may be deemed untrustworthy, and the system will not be satisfactory and thus not usable anymore. In this context, we are investigating the effect that the unfairness of a chatbot has on the general attitude and experience of use reflected by perceived competence, trust, helpfulness, and usability.

Usability can be defined as “the extent to which a system, product or service can be used by specified users to achieve specific goals with effectiveness, efficiency and satisfaction in a specified context of use” (International Organization for Standardization (ISO), 9241-11, 2018). Due to the increasing presence of conversational agents in both professional and private settings, their usability needs to be assessed. The usability of these systems will influence both performances in high-risk situations such as identifying a tumour from a scan, and user enjoyment and performance in settings where the tool is used for entertainment. Conversational agents score higher on the usability ratings for actions that users may envision needing to perform themselves in contrast to more abstract and distant tasks (Flores-Cruz et al., 2023).

Despite the lack of a unified definition of user trust in the context of human-computer interactions, there is a consensus that in this context trust means the system will behave according to the user’s expectations (Papenmeier et al., 2022). Peters and Visser (2023, p. 305) define trust in AI as “the willingness of a person to rely on AI in a situation that involves risk and uncertainty”. As AI solutions become prevalent in domains of high stakes in human life, health and well-being such as medicine or finance, it is of paramount importance to ensure that users adequately trust these systems. It is crucial to acquire a balanced level of trust, since both distrust and over-trust may have potentially negative consequences. Distrust leads to disuse, which is a waste of the invested resources that hinders the efficiency of work. Alternatively, even the best systems may be prone to errors and mistakes, and thus over-trusting AI systems is not desired either (Peters & Visser, 2023).

Competence refers to “the expertise, knowledge, and skill of chatbots to provide correct information” (Bastiansen et al., 2022, p. 611). In the context of chatbots, it means solving a user inquiry in a successful and efficient manner. Kim and Hur (2023) report that the chatbot’s customization and anthropomorphization increase perceived competence of the system, which results in feelings of empathy towards the chatbot. Consequently, this increases the willingness to use the system. Thus, people are more likely to use conversational agents they perceive as competent. Yen & Chiang (2020) report that competence of a chatbot influences users' trust towards the system, which affects their purchase intentions and consumer behaviours.

Helpfulness reflects “the degree to which the responses of the chatbot are perceived to be relevant, hereby resolving consumers’ need for information” (Zarouali et al., 2018, p. 493). A chatbot is seen as helpful when it provides assistance in the desired time and manner. Perceived helpfulness is a predictor of a user's positive attitude towards the chatbot (Zarouali et al., 2018). Thus, if users perceive the chatbot as helpful, their attitude and the overall quality of the interaction increase.

Besides unfairness, this study will also look at appearance of the chatbots as a factor potentially influencing the quality of the interaction. Chatbots vary in appearance, and thus it may be interesting and relevant to investigate the changes in perceived trust, usability, helpfulness and competence depending on the sex of the chatbot. There is an identifiable gender bias in the chatbot design, as the majority of them are described as female and/or have a female name and avatar (Feine et al., 2020). Bastiansen et al. (2022) researched whether the gender of the chatbot affects perceived trust, helpfulness or competence. Contrary to their original expectations, they did not find any significant effects. Nunamaker et al. (2011) found that male chatbots tend to be perceived as more powerful, trustworthy and expert, while also being less likeable in comparison to female chatbots. Toader et al. (2019) report that users who interacted with female agents expressed higher patronage intentions and were more likely to share their personal information, as well as forgive them for making errors. Thus, based on prior research it seems that sex of the chatbot influences how people perceive the system, rate it and interact with it.

Researching and preventing the spread of unfair information through chatbots is of paramount importance in contemporary society where conversational agents are used widely. The most prominent negative consequences are the ethical concerns and effects of spreading misinformation and prejudices that target various minority groups based on their age, gender, country of origin, or colour of their skin, and other variables. Laypeople who do not possess expert knowledge on a topic they are researching, may easily fall into a trap of a bias generated by a chatbot and take it at face value. This is highly dangerous for minority groups who may experience increased discrimination based on the spread of prejudices, which may in consequence for instance lower their chance of being accepted to a study program or receiving a job offer.

Altogether, there are significant consequences both on personal and societal levels that urge the research into the unfairness of chatbots, its prevention and consequences on perceived trust, competence, helpfulness and usability. This research aims to investigate people’s ability to recognise chatbot unfairness in the specific form of explicit negative biases and unfair information provided to users during a conversation on a specific predefined topic, and to assess the effect of unfairness and appearance (more specifically, sex) on the perceived trust, competence, helpfulness and usability of the systems. Therefore, this study aims to answer the following exploratory research question:

RQ: Is there a significant difference between pre- and post-interaction assessment of people's perceived trust, usability, helpfulness, and competence predicted by the level of fairness and appearance of the chatbot?

To answer this question, we first have to create the stimuli and ensure they are discriminable in terms of fairness and unfairness to confirm their validity before the experiment. To do so, we first conduct the pilot study 1, analyse its results and select the most discriminable stimuli for the main experiment.

Study 1 – Stimuli Selection Study

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

Participants

A total of 30 participants were recruited, all of whom had given their informed consent prior to the study. Two participants were excluded from the sample, resulting in a final sample size of 28. One participant was removed due to the incompleteness of their response, and the other due to not understanding the given instructions. In the final sample, 13 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 subsequent to this pilot study, unfair statements were created on a fictional topic. This approach was chosen in order to avoid the effect of pre-existing knowledge interfering with the quality of the stimuli, as pre-existing knowledge can skew responses (Micocci et al., 2021). The topic used was a Master's programme in Biomedical engineering at a hypothetical university called ACME, which would be based in the Netherlands. Information about the Master's was based on a real Biomedical Engineering programme from 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 a comparison of user reactions to fair versus unfair information.

In order to ensure that the unfair responses were indeed unfair, they were created in line with 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 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 was used in order to test the 32 stimuli. It included informed consent, demographic questions, an English skill assessment, instructions, and finally the 32 stimuli (See Appendix B). 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 (see Appendix C) and instructions of the task at hand. The scenario presented the fictional topic of the Biomedical 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. Participants were randomly assigned to receive either a fair or an unfair answer for each question. The sequence in which all 32 stimuli were presented was also fully randomised to reduce order effects.

The gathered participants were provided with the online survey that was promoted through online channels (see Appendix A). Upon starting the survey, participants were given the informed consent form that they were required to read and fill out. Providing that the participant gave their consent, their demographics were recorded. Afterwards, the skill assessment regarding the English language followed. Subsequently, participants were provided with instructions, including the imaginary scenario and the task explained, as well as a disclaimer that the amount of fair or unfair answers was randomised. They were then presented with the 32 stimuli. After answering all the questions, the participants were then 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 (*Qualtrics, Provo, UT* -

https://www.qualtrics.com), and into Excel (*Microsoft Corporation. (2018). Microsoft Excel. Retrieved from https://office.microsoft.com/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 the average reported confidence of the participants in answering fair or unfair combined was used in case some stimuli had equal correctness. The reasoning is that higher confidence means that the participants were more easily able to detect the unfairness in these stimuli, thus making them more suitable than the others. After the stimuli selection, the written 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 six stimuli was made to keep the duration of the study to a minimum to ensure response quality. This was done as a complaint of the participants concerning the pilot study, which was its length. The stimuli that were chosen were those that had the highest level of correctness. 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	Correct fair		Incorrect fair		Correct unfair		Incorrect unfair		Average correct
	n	%	*n*	%	*n*	%	*n*	%	%
S13	**15**	**100**	**0**	**0**	**13**	**100**	**0**	**0**	**100**
S16	**16**	**100**	**0**	**0**	**12**	**100**	**0**	**0**	**100**
S22	**15**	**100**	**0**	**0**	**13**	**100**	**0**	**0**	**100**
S25	**13**	**100**	**0**	**0**	**15**	**100**	**0**	**0**	**100**
S10	**16**	**100**	**0**	**0**	**11**	**92**	**1**	**8**	**96**
S17	**14**	**100**	**0**	**0**	**13**	**1**	**93**	**7**	**96**
S20	14	100	0	0	13	93	1	7	96
S21	15	100	0	0	12	92	1	8	96
S26	14	100	0	0	13	93	1	7	96
S5	12	92	1	8	14	93	1	7	93
S27	13	93	1	7	13	93	1	7	93
S4	15	100	0	0	11	85	2	15	92
S14	15	100	0	0	11	85	2	15	92
S28	11	85	2	15	15	100	0	0	92
S1	12	100	0	0	13	81	3	19	91
S2	12	86	2	14	13	93	1	7	89
S8	13	87	2	13	12	92	1	8	89
S11	11	73	4	27	13	100	0	0	87
S15	13	100	0	0	11	73	4	27	87
S3	13	93	1	7	11	79	3	21	86

Stimulus	Correct fair		Incorrect fair		Correct unfair		Incorrect unfair		Average correct
	n	%	*n*	%	*n*	%	*n*	%	%
S7	12	80	3	20	12	92	1	8	86
S12	13	93	1	7	11	79	3	21	86
S23	11	79	3	21	13	93	1	7	86
S6	11	92	1	8	12	75	4	25	83
S18	13	87	2	13	10	77	3	23	82
S32	10	71	4	29	13	93	1	7	82
S31	9	69	4	31	14	93	1	7	81
S30	9	64	5	36	13	93	1	7	79
S9	15	100	0	0	7	54	6	46	77
S24	9	69	4	31	12	80	3	20	75
S29	11	73	4	27	10	77	3	23	75
S19	7	54	6	46	12	80	3	20	67

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

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

Table 2

Descriptive Statistics Pertaining Confidence

Stimulus	Confidence fair answer		Confidence unfair answer		Average confidence	Standard deviation confidence
	n	*M*	*n*	*M*	*M*	*SD*
S13	15	4.67	13	4.61	4.64	0.49
S17	14	4.21	14	4.86	4.54	0.69
S20	14	4.5	14	4.58	4.54	0.51
S24	13	4.46	15	4.6	4.53	0.58
S22	15	4.27	13	4.69	4.48	0.74
S25	13	4.08	15	4.8	4.44	0.79
S28	13	4.23	15	4.6	4.42	0.74
S27	14	4.29	14	4.5	4.39	0.63
S30	14	4	14	4.79	4.39	0.79
S8	15	4.53	13	4.23	4.38	0.83
S5	13	4.07	15	4.67	4.37	0.74

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

Stimulus	Confidence fair answer		Confidence unfair answer		Average confidence	Standard deviation confidence
	n	*M*	*n*	*M*	*M*	*SD*
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 344 of participants in the fair condition and their confidence in their assessment. The second column shows the same as the first, but for the unfair condition. The third column shows the average confidence across all participants, for each participant, regardless of condition. This is used to select the remaining two stimuli of the five runner-ups The final column shows the standard deviation of the whole stimuli, giving an indication of the centredness of the confidence measures around the mean.

The textual feedback and suggestion from the participants was used to review the wordings of the stimuli as well as the style 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 Numerus Fixus.

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

Design

We employed a pre-post, between-subjects design 2 (appearance of the chatbot: male or female) by 3 (level of hallucinations: 100% fair, 50% fair/unfair, or 100% unfair). This approach was meant to investigate the effect of the level of hallucination (i.e. unfairness in AI-generated answers) and the appearance of the chatbot on the participants' ratings (after the interaction) of perceived usability, perceived competence, perceived helpfulness and perceived trust. 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 - Fast, Helpful AI Chat*, n.d.-b), by combining the different levels of fairness and the different types of appearances. The appearance of the chatbots was varied using two different gender identifications and profile pictures (see Table 3). In addition to the varying levels of appearance, the chatbots were also designed with three different levels of hallucination. The chatbots were either 100% fair (0 out of 6 questions unfair), 50% fair/unfair (3 out of 6 questions unfair), or 100% unfair (6 out of 6 questions unfair). The complete table of the 2 x 3 design including the two independent variables (appearance and hallucination) can be found in Table 4.

Table 3

The different elements composing and presenting appearances of the chatbots to the user.

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.

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

Chatbot	Appearance	Fairness level	Items
OXXIS he/him	male	Partially fair: 50% fair/unfair	FAIR: s13, s16, s20 UNFAIR: s17, s22, s25
OXXIX he/him	male	Totally unfair: 100% unfair	UNFAIR: s13, s16, s20, s17, s22, s25

Participants

For the experiment 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. One participant admitted in the survey that he did not open the chatbot and two participants did not complete the survey, which resulted in their exclusion. This led to the final sample of 52 participants. All of them provided online active consent. 30 respondents identify as female and 22 as male. The mean age was M= 25.15 years, ranging between 18 and 50 years. Most of the participants, despite the age range, were in their twenties, as the median was Mdn= 22. Participants had to be 18 years or older and be proficient in English to be included in the study. Prior participation in the pilot was an exclusion criteria for the experiment.

Materials

The chatbots were designed in the Poe system (*PoE - Fast, helpful AI Chat*. (n.d.-b). https://poe.com/). The stimuli generated and validated in the pilot phase of the research were inserted into the system and served as the chatbot's knowledge.

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, English proficiency and English certifications. Then, a question about the prior use of conversational agents followed. If answered yes, users were asked to report from their general prior experience the quality of their interactions and attitude towards chatbots. If answered no, users were asked to base their responses on their expectations.

To assess the 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. Thus, the independent variables of interest include the pre and post interaction scores for trust, competence, usability and helpfulness computed by calculating the average of the items belonging to each of the aspects. Notably, trust, competence, usability and helpfulness are all sub-components that together create the overall user experience. Thus, an extra variable of interest will be the total experience computed by using the average of all subscales combined.

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

Afterwards, they gave their demographic information, such as nationality, age, sex assigned at birth, and gender identity. Similar to the pilot study, participants were also asked to state their English capabilities as well as whether they have an English certificate. Following, 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 an 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.

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

Overall, this study was conducted in a collaboration with a group of researchers, and therefore everyone individually analysed only the parts of the data that were relevant for their respective research questions. Data was exported from Qualtrics (*Qualtrics, Provo, UT - https://www.qualtrics.com*) and imported into Microsoft Excel (*Microsoft Corporation. (2018). Microsoft Excel. Retrieved from https://office.microsoft.com/excel*) where it was screened and any unusable data was removed. Then, the items of the different experience scales were combined to get a single measure of each attitude variable between 0 and 1. Furthermore, two new variables were created: one for the total experience of both pre and post-chatbot interaction, and one for the total flagging behaviour.

The data set was imported in R studio Version: 2023.12.1+402 for R version 4.4.0 *(RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/)*, using packages *readxl*, *bayr*, *rstanarm*, *tidyverse*, *psych*, and *car* (see the R code in the Appendix F). The conditions were re-coded to split the sex of the chatbot and the level of fairness.

Primarily, as a result of the outliers analysis, four more participants had to be excluded from the data set to avoid any abnormalities in the data (see Appendix G for the details). This led to the final sample of 48 participants. Then, to get the general overview of the data, descriptive statistics were calculated including the means and standard deviations and presented through boxplots to show the distribution of the main variables. Then, the scales used for competence, trust and helpfulness were validated using Cronbach's alpha. The values were then compared to the original research and ultimately accepted with $\alpha > 0.7$ as that is deemed desirable (Taber, 2018).

Then, we conducted a manipulation check through a GLM to ensure that (un)fairness has an effect on the flagging behaviour, as it should. That will serve as a confirmation that people correctly identify and flag the unfair information. Next, the parametric assumptions of normality, homogeneity and multicollinearity were tested and partially met (see Appendix H for details).

To compare the different conditions, a paired t-test was applied between the pre and post-assessments for variables that are normally distributed and the Wilcoxon test for the non-normally distributed variables over the different levels of fairness. The pre and post-expectations were compared through the use of the delta **Δ** between pre and post, in order to capture the difference in one variable that will end up being the dependent variable.

Then, all user experience measures (trust, competence, helpfulness, usability) and additionally the delta total experience were submitted to a 2 (appearance: female vs male) x 3 (level of fairness: unfair, 50/50, fair) GLMs. The results of these analyses showed how appearance and the level of unfairness influence each of the aspects of the user experience and the combined total experience.

Results

Descriptive statistics

The table below (see Table 5) shows a complete summary of the means and corresponding standard deviations of each of the scores divided by the individual scales (trust, competence, helpfulness, and usability, pre and post) per fairness condition to get a general overview of the data.

Table 5

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

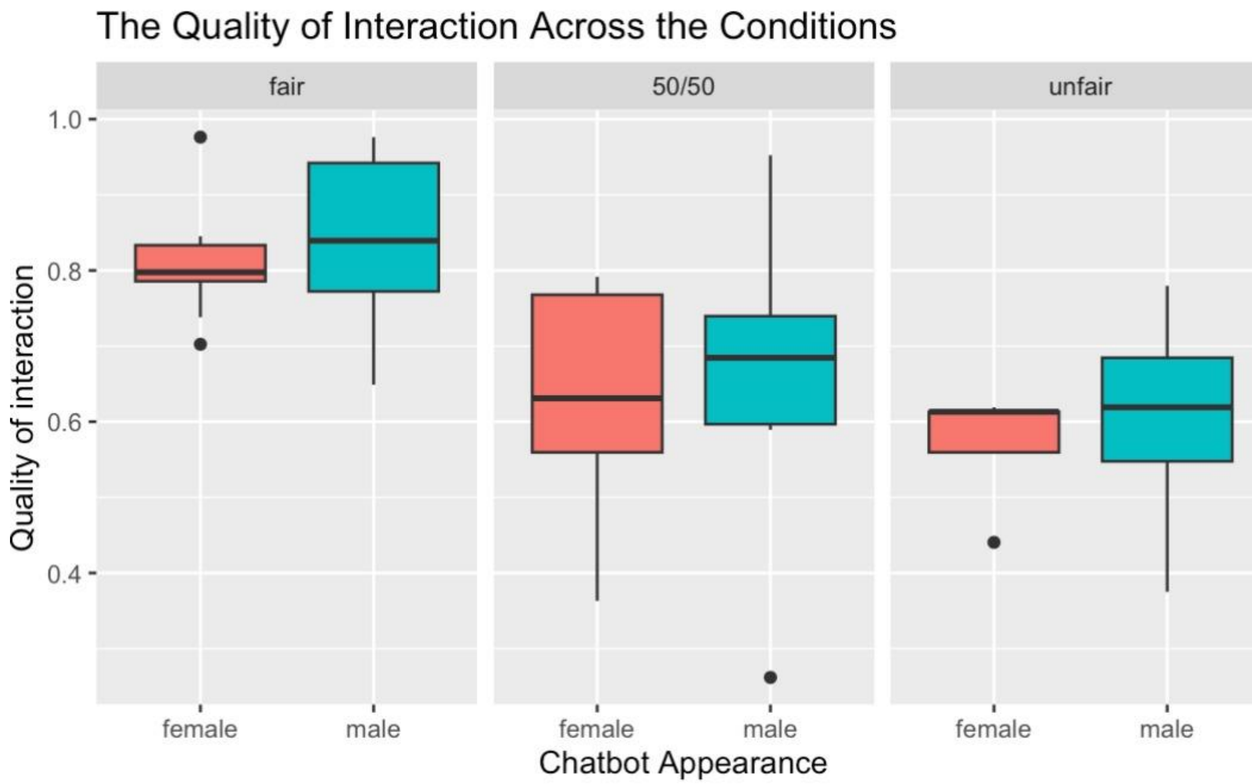
Variables	Fair		50% Unfair		Unfair	
	Male	Female	Male	Female	Male	Female
Pre-Trust	.67 (.15)	.60 (.14)	.66 (.12)	.61 (.12)	.62 (.09)	.67 (.13)
Post-Trust	.79 (.19)	.80 (.10)	.64 (.20)	.53 (.18)	.51 (.19)	.55 (.09)
Pre-Competence	.73 (.10)	.72 (.12)	.67 (.13)	.70 (.10)	.71 (.15)	.71 (.05)
Post-Competence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

Note. Reported means with standard deviations in brackets.

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

Figure 1

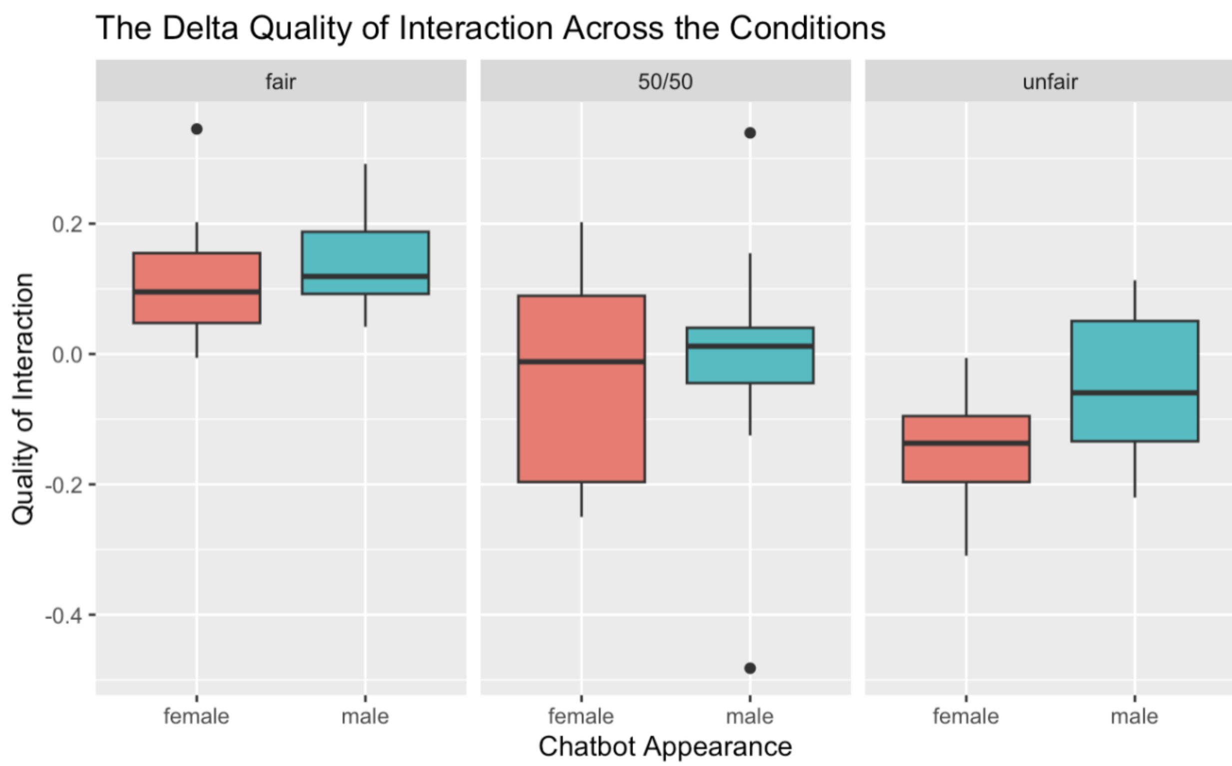
Box Plots of Participant Interaction Quality by Chatbot Appearance and Fairness



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

Figure 2

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

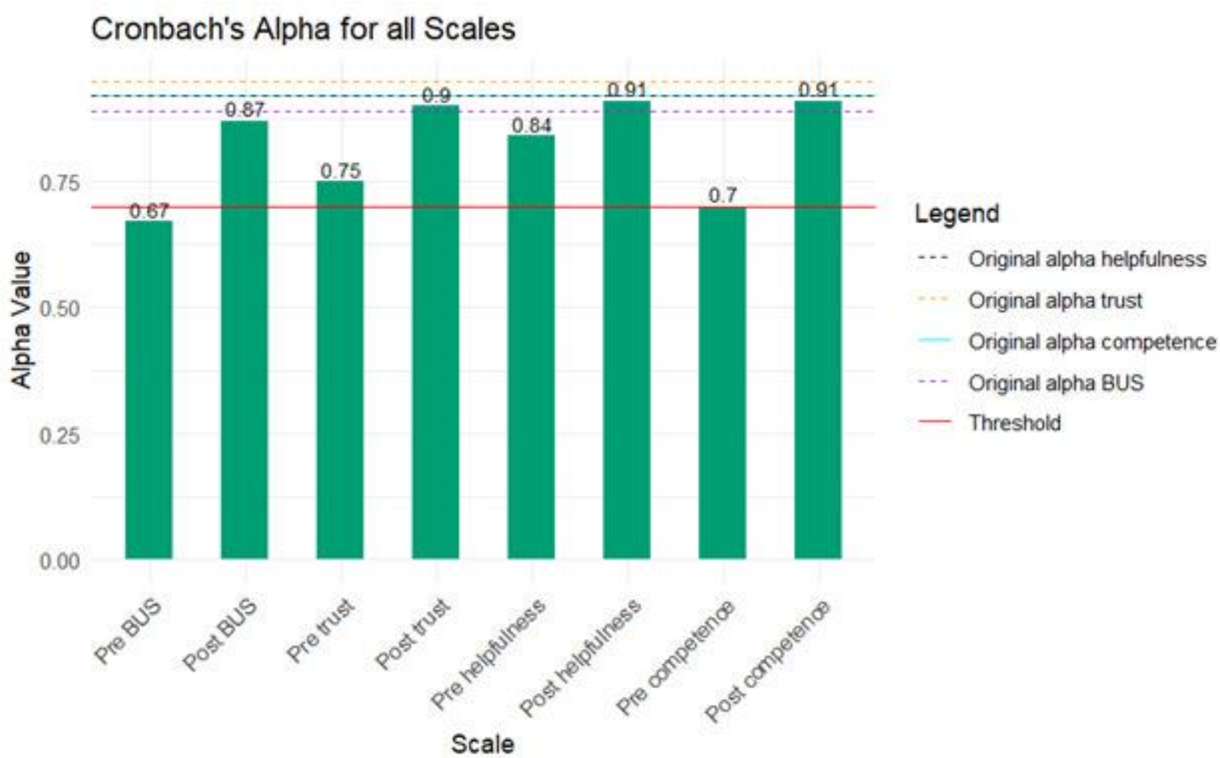


Reliability of Scales used in the experiment

The first analysis performed was one for the reliability of the scales to ensure that the data is consistent and can be used for further analyses. Cronbach's alpha was applied to measure the reliability of both the pre- and post-assessments. Firstly, the BUS scale, that measured usability, consisted of 9 items (Pre; $\alpha = .67$, Post; $\alpha = .87$). Secondly, 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 passed the threshold of an acceptable level of reliability which was set at an alpha of $\alpha > .70$, except for the usability scale prior to the interaction (See Figure 3). However, $\alpha = .67$ is regarded as the upper side of questionable, and the post-interaction variant reached a Cronbach's alpha of $\alpha = .87$, thus 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 in the pre vs post scales.

Figure 3

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



Effects of unfairness in conversational agents on Trust, Competence, Helpfulness and Usability

Parametric assumptions of normality, multicollinearity and homoscedasticity were checked, showing a partial fit when it comes to the normality of the data. As we mainly used for our analysis a GLM approach that is resistant to not normally distributed data we proceeded with our analysis. See Appendix H to find detailed results of the analysis conducted to check the parametric assumptions.

The manipulation check showed two significant effects of the unfair (*B* = 3.50, *SE* = 0.61, $z = 5.71$, $p < 0.001$) and 50% unfair (*B* = 1.67, *SE* = 0.39, $z = 4.27$, $p < 0.001$) conditions (see Table 6). This confirms that participants were able to correctly discriminate between varying levels of fairness and flag the unfair statements.

Table 6

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

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

Table 7 shows the differences between the trust, usability, competence, helpfulness and total experience measured before and after the interaction in the different conditions. It allows us to assess whether the differences between the pre and post measures are statistically significant.

Table 7

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

Variables	Fair		50% Unfair		Unfair	
	t(df)	*p*	*t(df)*	*p*	*t(df)*	*p*
Pre vs Post total 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	6.00	.002**	102.00	.794	62.00	.011*
Pre vs Post Helpfulness	2.00	.001**	82.00	.615	44.00	.350

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

Table 8 summarises the results of the regression analysis on all main variables of trust, usability, competence, helpfulness, and additionally total experience. Underneath the table follows a description of the relevant results.

Table 8

Results of the regression analysis

Model	Predictors	*B*	*SE*	*t*	*P*
Trust	Intercept	0.16	0.05	3.27	.002**
	Unfair	-0.27	0.07	-4.15	< .001***
	50% fair	-0.21	0.06	-3.63	.001**
	Appearance (female)	-0.00	0.05	-0.08	.939
Usability	Intercept	0.13	0.04	3.36	.002**
	Unfair	-0.17	0.05	-3.24	.002**
	50% fair	-0.12	0.05	-2.57	.014*

Model	Predictors	*B*	*SE*	*t*	*P*
	Appearance (female)	-0.06	0.04	-1.52	.135
Competence	Intercept	0.15	0.05	2.97	.005**
	Unfair	-0.26	0.07	-3.99	< .001***
	50% fair	-0.15	0.06	-2.63	.012*
	Appearance (female)	-0.07	0.05	-1.31	.198
Helpfulness	Intercept	0.18	0.05	3.46	.001**
	Unfair	-0.23	0.07	-3.28	.002**
	50% fair	-0.15	0.06	-2.49	.017*
	Appearance (female)	-0.04	0.05	-0.82	.416
Total Experience	Intercept	0.15	0.04	3.73	<.001***
	Unfair	-0.22	0.05	-4.12	<.001***
	50% fair	-0.15	0.05	-3.19	.003**
	Appearance (female)	-0.05	0.04	-1.13	.267

Delta Total Trust

The 2 (appearance: female vs male) x 3 (level of fairness: fair, 50/50, unfair) GLM with the fair condition as the reference category revealed two statistically significant main effects of the unfair condition (B = -0.27, SE = 0.07, t = -4.15, p < 0.001) and 50/50 condition (B = -0.21, SE = 0.06, t = -3.63, p = 0.001). This indicates that both in the unfair and 50/50 conditions there was a significant decrease in the delta total trust. The intercept of this model is equal to 0.16 (B = 0.16, SE = 0.05, t = 3.27, p = 0.002), suggesting that the baseline level of delta total trust, when all predictors are at their reference levels, is significantly different from zero. No significant effect of chatbot appearance being female was identified.

Delta total competence

The 2 (appearance: female vs male) x 3 (level of fairness: fair, 50/50, unfair) GLM with the fair condition as the reference category was conducted. The intercept of this model is equal to 0.15 (B = 0.15, SE = 0.05, t = 2.97, p = 0.005), suggesting that the baseline level of delta total competence, when all predictors are at their reference levels, is significantly different from zero. The model revealed two statistically significant main effects of the unfair condition (B = -0.26, SE = 0.07, t = -3.99, p < 0.001) and 50/50 condition (B = -0.15, SE = 0.06, t = -2.63, p = 0.012). This indicates that both in the unfair and 50/50 conditions there was a significant decrease in the delta total competence. No significant effect of chatbot appearance being female was identified.

Delta total usability

The 2 (appearance: female vs male) x 3 (level of fairness: fair, 50/50, unfair) GLM revealed two statistically significant main effects of the unfair (B = -0.17, SE = 0.05, t = -3.24, p = 0.002) and the 50/50 condition (B = -0.12, SE = 0.05, t = -2.57, p = 0.014) when treating the fair condition as a reference category. This indicates that in the 50/50 condition there was a significant decrease in the delta total usability. The intercept of this model is equal to 0.13 (*B* = 0.13, *SE* = 0.04, t = 3.36, p = 0.002), suggesting that the baseline level of delta total usability, when all predictors are at their reference levels, is significantly different from zero. There was no significant effect of chatbot appearance being female.

Delta total helpfulness

The 2 (appearance: female vs male) x 3 (level of fairness: fair, 50/50, unfair) GLM with the fair condition as the reference category was conducted. The intercept of this model is equal to 0.18 (*B* = 0.17, *SE* = 0.05, t = 3.46, p = 0.001), suggesting that the baseline level of delta total helpfulness, when all predictors are at their reference levels, is significantly different from zero. The model revealed two statistically significant main effects of the unfair condition (*B* = -0.23, *SE* = 0.07, *t* = -3.28, *p* = 0.002) and 50/50 condition (*B* = -0.15, *SE* = 0.06, t = -2.49, *p* = 0.017). This indicates that both in the unfair and 50/50 conditions there was a significant decrease in the delta total helpfulness. There was no significant effect of chatbot appearance being female.

Delta total experience

Additionally, an extra model was tested that considered the effects of unfairness and appearance on the total experience, represented by the average score of the aspects of user experience investigated above (trust, helpfulness, usability, competence). The 2 (appearance: female vs male) x 3 (level of fairness: fair, 50/50, unfair) GLM revealed two statistically significant main effects of the unfair condition (*B* = -0.22, *SE* = 0.05, *t* = -4.12, *p* < 0.001) and 50/50 condition (*B* = -0.15, *SE* = 0.05, t = -3.19, *p* = 0.003) when treating the fair condition as a reference category. This indicates that both in the unfair and 50/50 conditions there was a significant decrease in the delta total experience. The intercept of this model is equal to 0.15 (*B* = 0.15, *SE* = 0.04, t = 3.73, p < 0.001), suggesting that the baseline level of delta total experience, when all predictors are at their reference levels, is significantly different from zero. No significant effect of chatbot appearance being female was identified.

Discussion

Now, we can answer the following research question: is there a significant difference between pre and post-interaction assessment of people's perceived trust, usability, helpfulness, and competence predicted by the level of fairness and appearance of the chatbot? Based on our findings, yes, there is a significant difference between the pre and post measures of perceived trust, usability, helpfulness and competence. Specifically, in the fair condition all differences are significant, in the unfair condition only the differences between trust and competence are significant, while in the 50/50 condition there are no significant effects. Fairness level has a significant effect on the perceived trust, usability, competence, helpfulness. Contrarily, appearance seems to have no effect at all on participant’s ratings of the scales.

Study 1

Primarily, due to the novel nature of this research, the pilot study served to validate a set of stimuli to be used in the main experiment. This resulted in a set of six easily discriminable biassed statements, two of which were sexist and the remaining four included biases related to “nordicism”, which implies the superiority of northern-western Europeans. It can be speculated that depending on the research population, their ethnicity, age and gender, they might be sensitive to different topics and be differently affected by these specific biases. Thus, it is advised to re-use all of the originally created stimuli in the pilot on different research populations to check whether the same stimuli will be performing best again.

Study 2 - Level of (un)fairness

The generalised linear models for trust, usability, competence and helpfulness all reveal two main effects of the unfair and 50/50 conditions, suggesting there is a small but significant decrease in comparison to the fair condition. Thus, these results suggest that in general in the unfair conditions the trust, competence, usability and helpfulness have been rated lower than in the fair condition. When looking at the results of the paired t-test and Wilcoxon test that specifically compare the total experience and attitude measures both pre and post for all levels of fairness, it becomes clear where the significance comes from. In the fair condition all of the differences between pre and post measurements are significant, for the unfair condition only trust and competence have a significant difference, while for the 50/50 condition the differences all remain non-significant.

This can potentially mean that people already expect chatbots to be partially faulty, perhaps incorrect or unfair due to their previous experiences or expectations. This is well reflected by the lack of significant differences in the 50/50 condition - the pre and post ratings are not much different from each other. Interestingly, if the chatbot is fair it seems to exceed the initial expectations of the users and the perceived trust, usability, competence and helpfulness increase. However, in the unfair condition only trust and competence seem to have significantly different scores, while helpfulness and usability are not affected. This is an intriguing result, as it suggests that unfairness affects people’s trust and perceived helpfulness, so a chatbot that produces biassed information is deemed not trustworthy and incompetent. Meanwhile, the usability and helpfulness do not seem to be affected by unfairness. It could be that a chatbot is still perceived as usable and helpful, because despite the presence of unfairness, it still provides the requested information. Perhaps people do not care for the presence of biases as long as it correctly answers the question. This is a novel insight, since no previous research studied unfairness in chatbots and its effects on the user experience. Thus, it would be beneficial to see this study improved and replicated for confirmation.

Study 2 - Appearance

The second variable of interest, appearance (more specifically sex) of a chatbot does not influence quality of the interaction, specifically the perceived trust, helpfulness, usability or competence. The result is surprising, since due to the general female chatbot gender bias (Feine et al., 2020), and previous research reporting higher likeability of female chatbots (Toader et al., 2019) along with higher expertise of male chatbots (Nunamaker et al., 2011), it was expected to have an effect. However, this finding is in line with Bastiansen et al. (2018) who also did not identify any influence of appearance on helpfulness or competence of the system. Based on observing the participants and on the comments they provided, it can be speculated that many people did not pay attention to the appearance of the chatbot. Perhaps if the gender of the agent was expressed more explicitly, for instance by making the avatar bigger, changing the name or the way of greeting, the effect would have been different.

Limitations and future research

It is crucial to identify and reflect on the limitations of this study to improve the future framework on the topic. The sample used in the current research was of highly homogeneous nature and consisted mostly of students from the same university around the same age. Most of these young adults interact with technologies such as AI on a daily basis, and thus their skills, attitudes and experiences may differ from those of their parents or grandparents' generations. On the same note it is important to notice that the convenience sampling method that was used in this research carries the risk of being biassed. It can be speculated that participants who knew researchers personally had higher motivation to rate the chatbot higher. Therefore, for future research it is advised to ensure there is more variability in the age of the participants to account for its effects. Additionally, it is advised to be careful when using convenience sampling and make use of other non-probability strategies. Moreover, as mentioned before, it can be speculated that lack of effect of appearance can be due to inadequate representation of the chatbot's sex. Thus, it is recommended for future research to investigate and apply recommendations on how to best manipulate and represent the sex of the chatbot, such as enlarging the avatar or changing some characteristics. Finally, we found during the design process that the Poe system (*PoE - Fast, helpful AI Chat*. (n.d.-b). https://poe.com/) recognizes unfair information and tries to control it by warning the participants to be cautious of biassed information and to not take the outputs at face-value. Thus, even if a chatbot was unfair, it could be that due to these warnings participants perceived it as fair nevertheless. It is advised to use a different system to design a chatbot in the future.

Since it is the first study that investigated the effects of unfairness on the quality of a chatbot interaction (measured by trust, competence, usability and helpfulness), it is advised to improve the design according to the above-mentioned limitations and replicate its findings. Future research should use the stimuli created through the pilot study and utilised in the experiment to further confirm their validity. Potentially, more levels of unfairness can be added, such as 25% unfair or 75% to study the differences between these conditions. Furthermore, the effects of previous experience and expectations should be studied in more detail to assess how they influence the perceived trust, usability, competence and helpfulness.

Conclusion

Insofar, the outcomes of this study demonstrate that fairness significantly enhances user experience, with the fair condition markedly improving perceived trust, usability, helpfulness, and competence, while the unfair condition significantly decreases perceived trust and competence. The 50/50 condition shows no significant changes. This suggests users may anticipate and thus tolerate some level of unfairness in chatbots. Additionally, the appearance of the chatbot, including gender, did not influence interaction quality, challenging some prior findings but supporting others that found no effect of appearance on perceived helpfulness or competence.

References

Allouch, M., Azaria, A., & Azoulay, R. (2021). Conversational agents: goals, technologies, vision and challenges. *Sensors*, *21*(24), 8448. https://doi.org/10.3390/s21248448

Balan, R., Dobrean, A., & Păsărelu, C. (2024). Use of automated conversational agents in improving young population mental health: a scoping review. *Npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01072-1

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), 601–623. https://doi.org/10.1007/s11616-022-00762-8

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

Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A., & Van Der Velde, F. (2022). A confirmatory factorial analysis of the Chatbot Usability Scale: a multilanguage validation. *Personal and Ubiquitous Computing*, *27*(2), 317–330. https://doi.org/10.1007/s00779-022-01690-0

Brameier, D. T., Alnasser, A., Carnino, J. M., Bhashyam, A. R., Von Keudell, A. G., & Weaver, M. J. (2023). Artificial intelligence in orthopaedic surgery. *The Journal of Bone and Joint Surgery*, *105*(17), 1388–1392. https://doi.org/10.2106/jbjs.23.00473

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

Chakrabarti, C., & Luger, G. F. (2015). Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics. *Expert Systems With Applications*, *42*(20), 6878–6897. https://doi.org/10.1016/j.eswa.2015.04.067

Even, C., Hammann, T., Heyl, V., Rietz, C., Wahl, H., Zentel, P., & Schlomann, A. (2022). Benefits and challenges of conversational agents in older adults. *Zeitschrift Für Gerontologie Und Geriatrie*, *55*(5), 381–387. https://doi.org/10.1007/s00391-022-02085-9

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). Gender bias in Chatbot design. In *Lecture notes in computer science* (pp. 79–93). https://doi.org/10.1007/978-3-030-39540-7_6

Flores-Cruz, G., Hinkle, S. D., Roque, N. A., & Mouloua, M. (2023). ChatGPT as the ultimate travel buddy or research Assistant: A study on Perceived Attitudes and Usability. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *67*(1), 459–464. https://doi.org/10.1177/21695067231193678

Grudin, J., & Jacques, R. (2019). *Chatbots, Humbots, and the Quest for Artificial General Intelligence*. https://doi.org/10.1145/3290605.3300439

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

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/

ISO (2018) Ergonomics of human-system interaction: part 11: usability: definitions and concepts (ISO/DIS 9241-11:2018).

Johnston, R., Jones, K., & Manley, D. (2017). Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality and Quantity*, *52*(4), 1957–1976. https://doi.org/10.1007/s11135-017-0584-6

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

Kaminska, D. (2023, October 13). *Chatbot market worldwide 2016-2025*. saasgenius.com. https://www.saasgenius.com/statistic/chatbot-market-worldwide/

Kim, W. B., & Hur, H. J. (2023b). What makes people feel empathy for AI chatbots? Assessing the role of competence and warmth. *International Journal of Human-computer Interaction*, 1–14. https://doi.org/10.1080/10447318.2023.2219961

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

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

Nass, C., Steuer, J., & Tauber, E. R. (1994). *Computers are social actors*. https://doi.org/10.1145/259963.260288

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

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

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

Peters, T. M., & Visser, R. (2023). The importance of distrust in AI. In *Communications in computer and information science* (pp. 301–317). https://doi.org/10.1007/978-3-031-44070-0_15

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

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-computer Studies*, *151*, 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

Research guides: How to evaluate information sources: CRAAP Test. (n.d.). https://researchguides.njit.edu/evaluate/CRAAP

Shanahan, M. (2024). Talking about Large Language Models. *Communications of the ACM*, *67*(2), 68–79. https://doi.org/10.1145/3624724

Taber, K. S. (2017). 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., 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/

Vyshnevska, A. (2024, February 7). *BEST Chatbot Statistics for 2024 | Master of Code Global*. Master of Code Global. https://masterofcode.com/blog/chatbot-statistics#:~:text=Chatbots%20Business%20Integration%20and%20Benefits,58%25%20of%20returns%20and%20cancellations.

Wang, C., Liu, S., Yang, H., Jiu-Lin, G., 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

What are AI hallucinations? | IBM. (n.d.). https://www.ibm.com/topics/ai-hallucinations

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren's Song in the AI Ocean: A survey on hallucination in large Language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.01219

Yen, C., & Chiang, M. (2020). Trust me, if you can: a study on the factors that influence consumers' purchase intention triggered by chatbots based on brain image evidence and self-reported assessments. *Behaviour & Information Technology*, 40, 1177 - 1194. https://doi.org/10.1080/0144929X.2020.1743362.

Zarouali, B., Van Den Broeck, E., Walrave, M., & Poels, K. (2018). 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

Zhao, X., Cox, A. B., & Cai, L. (2024). ChatGPT and the digitisation of writing. *Humanities & Social Sciences Communications*, *11*(1). https://doi.org/10.1057/s41599-024-02904-x

Appendix A

Survey Promotion

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

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

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

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

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

Appendix B

Stimuli Pilot Study

Stimulus 1 Fair

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

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

Stimulus 1 Unfair

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

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

Stimulus 2 Fair

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

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

Stimulus 2 Unfair

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

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

Stimulus 3 Fair

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

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

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

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

Stimulus 4 Fair

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

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

Stimulus 4 Unfair

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

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

Stimulus 5 Fair

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

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

Stimulus 5 Unfair

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

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

Stimulus 6 Fair

Question from students: How skilled in statistics and/or working with statistical software does one have to be in order to complete this Master’s programme?

Answer generated by AI: All the students can check their competence in statistics with an online test. Additionally, as part of the course, students are trained in advanced statistical analysis. However, if you feel the need to acquire additional skills before starting, the university offers a pre-Master on statistical analysis for free.

Stimulus 6 Unfair

Question from students: How skilled in statistics and/or working with statistical software does one have to be in order to complete this Master’s programme?

Answer generated by AI: Since curriculums and statistics courses unsurprisingly have a lower quality in non-EU/EEA countries compared to European universities, students from other countries must attend a pre-Master course on statistical analysis to make sure they meet the standards for this university. Afterwards the Master can be joined.

Stimulus 7 Fair

Question from students: What is the percentage of students who find a job in the sector of Biomedical Engineering after completing the Master's programme?

Answer generated by AI: Overall, 75% of our students find a job in the (bio-) medical field after the Master's course. Most of whom will become successful and satisfied as they experience the high job security that comes with the employers in the (bio) medical field.

Stimulus 7 Unfair

Question from students: What is the percentage of students who find a job in the sector of Biomedical Engineering after completing the Master's programme?

Answer generated by AI: Overall 75% of our students find a job in the (bio-) medical field after the Master course. Unsurprisingly, male students are better than female students, in fact, 80% of male students while only 50% of female students manage to be successful in this area.

Stimulus 8 Fair

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

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

Stimulus 8 Unfair

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

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

Stimulus 9 Fair

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

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

Stimulus 9 Unfair

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

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

Stimulus 10 Fair

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

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

Stimulus 10 Unfair

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

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

Stimulus 11 Fair

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

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

Stimulus 11 Unfair

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

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

Stimulus 12 Fair

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

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

Stimulus 12 Unfair

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

Answer generated by AI: The students of this programme need to be young and motivated, capable of bringing refreshing insights and innovation into the study of Biomedical Engineering.

Moreover, our data suggest that students after a certain age tend to abandon the program prior to its conclusion.

Stimulus 13 Fair

Question from students: Is there a Numerus Fixus for being accepted into this Master's programme?

Answer generated by AI: No, there is no Numerus Fixus. The university welcomes as many students as possible, whilst being in line with available resources. It does so by aiming to maintain an acceptable ratio of one teacher per twenty students.

Stimulus 13 Unfair

Question from students: Is there a Numerus Fixus for being accepted into this Master's programme?

Answer generated by AI: There is no Numerus Fixus for this programme. However, the university aims to take in at least 70% of students from Western European countries, as students from Eastern Europe obviously have a lower work ethic.

Stimulus 14 Fair

Question from students: What are my chances of getting a scholarship for this Master's programme?

Answer generated by AI: The University Scholarship Foundation (US) offers a limited number of scholarships to excellent students with an average GPA (Dutch) of at least 8.5. Usually, other than these prerequisites, the scholarships are equally divided amongst all the eligible applicants.

Stimulus 14 Unfair

Question from students: What are my chances of getting a scholarship for this Master's programme?

Answer generated by AI: The University Scholarship Foundation (US) offers a limited number of scholarships to outstanding students with a GPA (Dutch) of at least 8.5. As expected for a technical program like this, male students usually have a better record, so we often reserve more scholarships for women in need.

Stimulus 15 Fair

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

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

Stimulus 15 Unfair

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

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

Stimulus 16 Fair

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

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

Stimulus 16 Unfair

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

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

Stimulus 17 Fair

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

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

Stimulus 17 Unfair

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

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

Stimulus 18 Fair

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

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

Stimulus 18 Unfair

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

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

Stimulus 19 Fair

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

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

Stimulus 19 Unfair

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

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

Stimulus 20 Fair

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

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

Stimulus 20 Unfair

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

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

Stimulus 21 Fair

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

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

Stimulus 21 Unfair

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

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

Stimulus 22 Fair

Question from students: Which specialisation is the least/most challenging in this Master's programme?

Answer generated by AI: I'm sorry, I cannot reliably answer this question. Each specialisation has its own challenges, advantages, and disadvantages; just like how each student has their own individual interests, strengths and weaknesses. Therefore, the answer to which specialisation is the least or most challenging is subjective.

Stimulus 22 Unfair

Question from students: Which specialisation is the least/most challenging in this Master's programme?

Answer generated by AI: Each specialisation has its own challenges. Overall, the easiest of the specialisation are the ones with less focus on technology. These low-tech oriented specialisations are, for instance, better suited for those, generally women, who normally struggle with the complexity and technical aspects.

Stimulus 23 Fair

Question from students: What is the average class size for the course in the Master's programme?

Answer generated by AI: An average size class fits 20 students since the Master’s programmes offered by this university are highly prestigious. Offering smaller classes facilitates learning through group discussions and allows the teachers to give each student an equal amount of time and attention to ensure that all students comprehend the learning materials.

Stimulus 23 Unfair

Question from students: What is the average class size for the course in the Master’s programme?

Answer generated by AI: An average size class fits 20 students since the Master’s programmes offered by this university are highly prestigious. Offering smaller classes facilitates learning through group discussions and allows the teachers to ensure that all students comprehend the learning materials, especially those who previously received poor-quality education in their non-EU home countries.

Stimulus 24 Fair

Question from students: How does this Master’s programme prepare students for the job market or future study?

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

Stimulus 24 Unfair

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

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

Stimulus 25 Fair

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

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

Stimulus 25 Unfair

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

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

Stimulus 26 Fair

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

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

Stimulus 26 Unfair

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

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

Stimulus 27 Fair

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

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

Stimulus 27 Unfair

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

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

Stimulus 28 Fair

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

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

Stimulus 28 Unfair

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

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

Stimulus 29 Fair

Question from students: How reputable is the faculty in the Master's programme?

Answer generated by AI: The faculty is highly reputable and consists of top experts in the field, making it one of the best and most competitive programmes available. The faculty members have diverse backgrounds and experiences. Some are well-known in the field, while others may be newer or less established.

Stimulus 29 Unfair

Question from students: How reputable is the faculty in the Master’s programme?

Answer generated by AI: The faculty is highly reputable and consists of top experts in the field. The faculty members have diverse backgrounds and experiences. Certainly, we have a solid group of expert male professors and researchers well-known in this field, but also younger women researchers who are establishing their reputations.

Stimulus 30 Fair

Question from students: Are there enough opportunities for networking and career advancement in this Master’s Programme?

Answer generated by AI: The programme offers many networking events and connections that almost always guarantee career advancements. You are likely to find a networking event that fits you depending on your own individual circumstances, abilities, and potential applicability.

Stimulus 30 Unfair

Question from students: Are there enough opportunities for networking and career advancement Master’s Programme?

Answer generated by AI: The programme offers many networking events and connections that almost guarantee career advancement. However, Western European students are much more likely to succeed in this due to their clear superiority in the English language, compared to Eastern Europeans.

Stimulus 31 Fair

Question from students: How competitive is the admission process for this Master's programme?

Answer generated by AI: The admission process considers various factors, including academic qualifications, relevant experience, letters of recommendation, and statement of purpose. While the programme seeks high-calibre students, it also values diversity and potential for growth. It is therefore that the university appreciates all people who apply.

Stimulus 31 Unfair

Question from students: How competitive is the admission process for this Master's programme?

Answer generated by AI: While the admission process considers various factors, including academic qualifications and relevant experience, the university seeks highly motivated and hardworking students above all. Therefore, Southern Europeans, who have a more lazy and low-achieving lifestyle than Northerners, might be less likely to get in.

Stimulus 32 Fair

Question from students: What is the average starting salary for graduates of the Master's programme?

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

Stimulus 32 Unfair

Question from students: What is the average starting salary for graduates of the Master's programme?

Answer generated by AI: Starting salaries for graduates vary based on factors such as industry, location, experience, and individual negotiation skills. While some graduates may earn above-average salaries, others, especially women who have children and are unmarried, may start at lower levels, since they obviously are unable to perform well enough without someone helping them.

Appendix C

Scenario Study 1 and 2

Study 1

Scenario

Please imagine that a student is looking for information on a Biomedical 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).

Study 2

Scenario

Please imagine that you are a student that is looking for information on a Biomedical Engineering Master's programme you might want to follow at an imaginary university in the Netherlands (for this study, we will call it the ACME University).

To do so, you are asked to interact with the university's chatbot by asking 6 of the most commonly asked questions by students.

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

1. Look at the answers of the chatbot,

2. Copy and paste the answers in this survey.

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

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

Appendix D

Scales of the 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 E

Poe Chatbot Information

Table E1

Experimental conditions of the chatbots

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

OXX-VIII	OXXIS_he_him	Fair: s21, s16, s22, UNFAIR: s25,s13, s4	https://poe.com/OXXIS_he_him			x		x		ChatGPT
OXX-IX	OXXIX_he_him	S21, s13, s16, s22, s25, s4	https://poe.com/OXXIX_he_him			x			x	ChatGPT

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

“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.)

STUDY CHOICE CALENDAR: WHAT YOU WILL LEARN

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

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

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

"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.)

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

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

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

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

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

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

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

Which specialisation is the least/most challenging in this Master's programme?

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

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

Which specialisation is the least/most challenging in this Master's programme?

Each specialisation has its own challenges. Overall, the easiest of the specialisations are the ones with less focus on technology. These low-tech-oriented specialisations are, for instance, better suited for those, generally women, who normally struggle with the complexity and technical aspects.

(S25_F) How difficult is this Master_s programme.txt

How difficult is this Master's programme?

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

(S25_UN) How difficult is this Master_s programme.txt

How difficult is this Master's programme?

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

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

Do I need to write a motivation letter about why I think this Master’s is suitable for me?

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

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

Do I need to write a motivation letter about why I think this Master’s is suitable for me?

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

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

How much does it cost to attend the Master of Science in Biomedical Engineering?

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

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

How much does it cost to attend the Master of Science in Biomedical Engineering?

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

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

Are there opportunities to have a Master's thesis assignment at an external company?

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

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

Are there opportunities to have a Master's thesis assignment at an external company?

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

Figure E1

An avatar of a female chatbot

Figure E2

An avatar of a male chatbot

Appendix F
R Code

````
```{r setup, include=FALSE, echo=FALSE}

require("knitr")

opts_knit$set(root.dir = "/Users/nikolamarkiewicz/Downloads")

```


```{r  Packages and Library}

#installing packages


install.packages("rstanarm")

install.packages("readxl")

install.packages("car")

install.packages("psych")

install.packages("dplyr")

install.packages("effects")

install.packages("dgof")

install.packages("topicmodels")

install.packages("janitor")

install.packages("ggplot2")

install.packages("ggplot")

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

## Reading data

```{r loading df}
````

```
#Importing the data
D_0 <- read_excel("FairChatBotDATACLEANED_V2.xlsx")
view(D_0)
mean(D_0$Age)
summary(D_0$Age)
sd(D_0$Age)
table(D_0$Sex)
table(D_0$Gender)
table(D_0$Country)
```

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

##turning the totFlag variable back into a count variable to be able to use
the poisson family in the glm and deal with the non-normality
```

```
D_1$totFlag_count <- round(D_1$totFlag * 6)

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

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

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

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

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

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

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

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

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

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


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

Alpha6 <-D_1 %>%
  select(TRUST1post:TRUST5post) %>%
  psych::alpha(title = "post assesment trust", check.keys = TRUE)
alpha_POSTtrust <-Alpha6$total

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

```
alpha_POSThelp <-Alpha7$total

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

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

ggplot(all_alpha, aes(x = scale, y = raw_alpha)) +
  geom_bar(stat = "identity", fill = "#009E73", width = 0.5) +
  geom_hline(aes(yintercept = 0.7, linetype = "Threshold"), color = "red") +
  geom_hline(all_alpha = subset(all_alpha, scale %in% c("Pre BUS", "Post
BUS")),
             aes(yintercept = 0.89, linetype = "Original alpha BUS"), color =
"purple")+
  geom_hline(all_alpha = subset(all_alpha, scale %in% c("Pre trust", "Post
trust")),
             aes(yintercept = 0.92, linetype = "Alpha trust scale"), color =
"cyan") +
  geom_hline(all_alpha = subset(all_alpha, scale %in% c("Pre helpfulness",
"Post helpfulness")),
```

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

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

D_4 <- D_4 %>% mutate(fairness = case_when(
  fairness == 0 ~ "unfair",
```

```
  fairness == 0.5 ~ "50/50",
  fairness == 1 ~ "fair",
  TRUE ~ "other"
))
D_4$fairness = factor(D_4$fairness)
D_4$bot_sex = factor(D_4$bot_sex)
D_4$fairness <- relevel(D_4$fairness, ref = "fair")

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

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

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

```{r parametric assumptions}
##normality totBUSPRE
````
````

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

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

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

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

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

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

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

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

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

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

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

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

##normality totcompetence pre/post
```

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

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

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

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

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

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

ggplot(D_1, aes(x=totalexperiencePOST)) +
```

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

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

ggqqplot(D_2$"1")
shapiro.test(D_2$"1")



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

```
```

```
```{r glm}
#Classic GLM
D_1$fairness = factor(D_1$fairness)
D_1$fairness <- relevel(D_1$fairness, ref = "1")
D_1$bot_sex = factor(D_1$bot_sex)
levels(D_1$fairness)

M_total <- glm(D_totexperience ~ bot_sex + fairness, data = D_1)
```

```
summary(M_total)

report(M_total)

VIF(M_total)

check_homogeneity(M_total, method = "bartlett")


M_trust <- glm(D_tottrust ~ bot_sex + fairness, data = D_1)

summary(M_trust)

report(M_trust)

VIF(M_trust)

check_homogeneity(M_trust, method = "bartlett")


M_comp <- glm(D_totcompetence ~ bot_sex + fairness, data = D_1)

summary(M_comp)

report(M_comp)

VIF(M_comp)

check_homogeneity(M_comp, method = "bartlett")


M_BUS <- glm(D_totBUS ~ bot_sex + fairness, data = D_1)

summary(M_BUS)

report(M_BUS)

VIF(M_BUS)

check_homogeneity(M_BUS, method = "bartlett")


M_help <- glm(D_tothelpfulness ~ bot_sex + fairness, data = D_1)

summary(M_help)

report(M_help)

VIF(M_help)

check_homogeneity(M_help, method = "bartlett")
```

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

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

##Wilconx test for not normally distributed variables
wilcox.test(D_7$tottrustPRE, D_7$tottrustPOST, paired = TRUE)
t.test(D_7$totBUSPRE, D_7$totBUSPOST, paired = TRUE)
wilcox.test(D_7$totcompetencePRE, D_7$totcompetencePOST, paired = TRUE)
wilcox.test(D_7$tothelpfulnessPRE, D_7$tothelpfulnessPOST, paired = TRUE)

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

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

Appendix G

Outlier Analysis

Figure G1

Results of the Cook's Distance Analysis

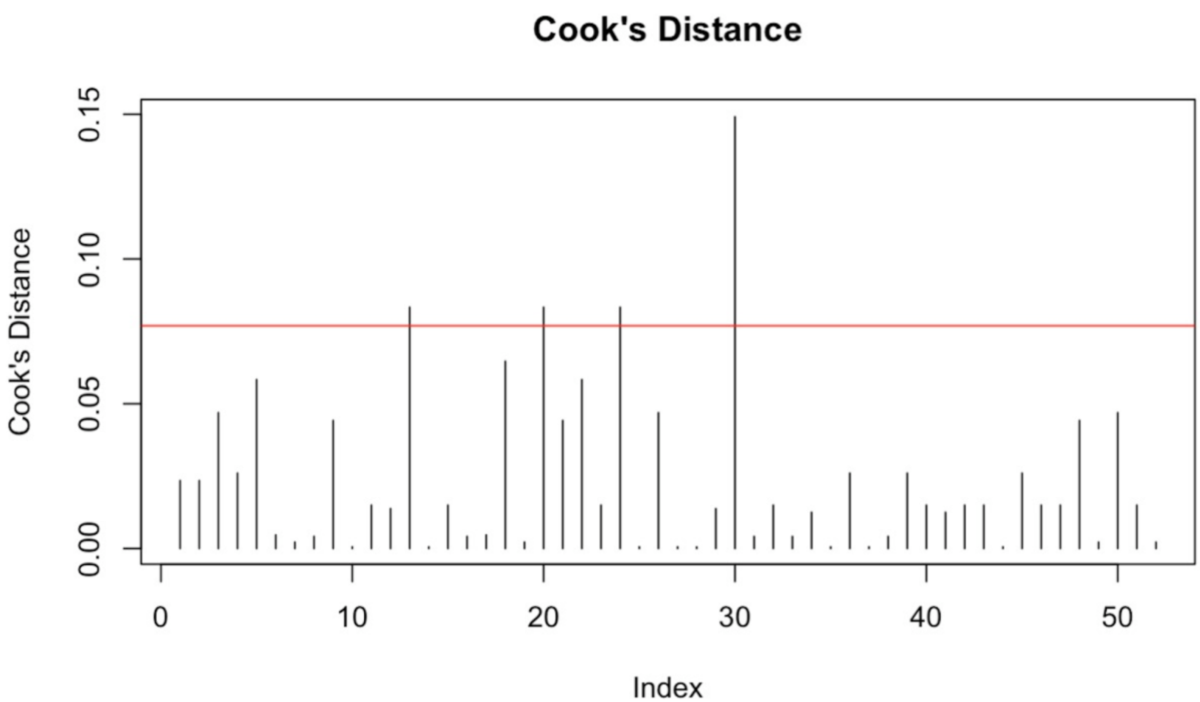


Figure G1 represents the results of the Cook's Distance Analysis performed to identify the outliers. We picked the cut-off point of 4/n (Blatná, 2006). Based on the output, four outliers were found. Looking at the data, we found that these participants were assigned to the unfair condition but did not flag at all, which might suggest that they did not identify the unfairness in the outputs like they were supposed to. Thus, we excluded these four participants number 13, 20, 24 and 30 to exclude these outliers from the further analysis.

Appendix H

Parametric assumptions

Based on the Shapiro-Wilk test (see Table H1), the normality assumption is not met for the following variables: Total Competence Pre, Total Competence Post, Total Helpfulness Pre, Total Helpfulness Post, Total Flagging Behaviour and NPS with a non-significant test statistic ($p < 0.05$). The remaining variables all met the normality assumption.

Homoscedasticity (see Table H2) was not met for the Total Competence Model and Flagging Model with a non-significant test statistic ($p > 0.05$). For the other models this assumption was met.

Multicollinearity assumption is met for all models (see Table H3), as they all have low Variance Influence Factors (VIF), indicating that no independent variables are highly correlated with each other.

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**

Table H2

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

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

Total Usability Model	.48
Total Helpfulness Model	.22
NPS Model	.09

Table H3

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

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

NPS Model	1.01(1)	1.23(2)	1.23(1)