

**Quality of Interaction With Chatbots as a Function of Fairness, Appearance and
Congruence With Participants' Declared Sex**

Lucas J. Assen

Department of Psychology, University of Twente

Faculty of Behavioural Sciences, Management and Social Sciences

202000384: BSc Thesis PSY

Supervisor: Dr. Simone Borsci, Second assessor: Dr. Maryam Amir Haeri

July 05, 2024

Abstract

The availability of knowledge through the information age is of great benefit to anyone. However, it also enables the spread of unfair information i.e., biased and incorrect information which, through filter bubbles, can lead to polarisation. It is known that unfair information may spread through artificial intelligence-driven chatbots (CA) due to the hallucination effect. However, no previous work has been done to understand its influence on the human-CA interaction, this is the current study's novelty. The present study used factors known to play a role in the quality of interaction with CAs to test the influence (trust, usability, helpfulness, and competence). Separately, the influence of the CA's appearance and congruence were investigated. This led to two problems that were empirically explored; i) whether the CA's appearance and congruence influence the flagging of information; and ii) whether the difference between the user's expected and experienced quality of interaction differs across levels of (un)fairness and whether the CA's appearance and congruence have an influence. Due to the novelty of the research area, two studies were done. The first focussed on creating a set of clearly unfair stimuli. This resulted in six stimuli that were selected as best performers. The second study focussed on the experimental assessment of the human-CA interactions. Here, a pre-post, between-subjects design was used that was 2 (appearance of the CA: male or female) by 3 (level of fairness: Completely fair, 50% unfair, or completely unfair), resulting in the creation of six CA. The results showed a difference in the expected and experienced quality of interaction across the completely fair and unfair conditions. The difference was due to an increase in the quality of interaction in the completely fair condition and due to a slight decrease in the completely unfair condition, whilst no difference was found in the 50% unfair condition. The results suggest that people expect CAs to be somewhat unfair. Lastly, no differences due to appearance nor congruence were found on either the flagging of information or the quality of interaction adding evidence to the debate as to whether such effects exist. In the end, additional recommendations for future works are given.

Keywords: Fairness, Appearance, Congruence, Human-CA interactions, Stimuli creation

Table of Contents

Abstract	2
Introduction	5
Study 1 – Stimuli Selection Study	12
Methods of Study 1	12
Participants	12
Materials	13
Procedure	14
Data Analysis	15
Results of Study 1	15
Study 2 – Experimental Assessment of Interactions with (Un)fair CAs	19
Methods of Study 2	19
Design	19
Experimental conditions	19
Participants	21
Materials	21
Procedure	23
Data Analysis	24
Results of Study 2	27
Descriptive statistics for the main variables of the data	27
Reliability analysis of the quality of interaction subcomponents	31
Manipulation check for the levels of (un)fairness	33
Exploration and investigation of the research questions	33
Discussion	37
The Novelty of Study 1	40
Limitations of the Study and Recommendations for Future Works	40
References	44
Appendix A: Survey Promotion Study 1	54
Appendix B: Generated Stimuli Study 1	55
Appendix C: Introduction, Scenarios, Instructions and Disclaimers of Study 1 and 2	73
Appendix D: Scales Study 2	78
Appendix E: Poe Chatbot Information	86
Appendix F: R Code	92

Appendix G: Outlier Analysis & Parametric Assumptions..... 104

Introduction

Nowadays people are continuously exposed to information about anything at any time, and anywhere (de Vreese & Neijens, 2016). This came to be due to the development of pervasive information technologies (IT) e.g., the invention of computers, which led to the inception of the internet that in turn made possible the birth of the current hyperconnected society through social media (de Vreese & Neijens, 2016). Whilst it greatly improved access to information, the innovations created by this information age also bring along downsides, it makes people more exposed to incorrect and untruthful information in several ways (Roetzel, 2019; Muhammed & Mathew, 2022). A first example is fake news which due to novelty and shock titles spreads faster than truthful news (Muhammed & Mathew, 2022). This incentivises people to spread fake news; it becomes a money-making machine, for people, because of the clicks and views it gets (Cooke, 2017). A second downside is the use of AI curation algorithms, which match content to the user's taste (Modgil et al., 2021; Valentino et al., 2009). AI curation algorithms can lead to polarisation through a reduction of passive exposure to a broad selection of content. It also leads to the satisfaction of confirmation biases meaning they show users proof of something they already believe in regardless of factual correctness (Modgil et al., 2021; Valentino et al., 2009). Following this, according to Berkman and Katona (2020), the consequences of AI curation algorithms can lead to a 'filter bubble' in which a person receives and consumes information that can be biased, wrong, incorrect and fake as long as it caters to the individual's interest. Filter bubbles are a detriment to the user, they lead to the reinforcement of a person's biases, whilst on a broader scale, they lead to polarisation (Dingler et al., 2018; Zhang et al., 2024).

To understand how people end up in such a 'filter bubble' attention is given to how people perceive biased, wrong, and fake information, which it is made up of, this will be referred to as unfair information as an overarching term moving forward. The first reason why people end up in a 'filter bubble' is because, when presented with both fair and unfair information, people are more interested in unfair information, this is often due to the shock value of unfair information (Muhammed & Mathew, 2022; Streicher et al., 2011). As suggested by Castillo et al. (2011), unfair information is less often fact-checked than fair information, as the negative sentiment related to unfair information tends to give it a sense of credibility. This shows that people are susceptible to unfair information which extends to landing in a 'filter bubble. The issue of filter bubbles is magnified when realising that people have a hard time correcting their beliefs even after being made aware that the information is unfair. Here, an especially large difference is seen between those with low and high cognitive

abilities (de Keersmaecker & Roets, 2017). On top of that, even when people acknowledge and agree that the information is unfair, they tend to continue using it in their reasoning, which is coined as the continued influence effect (Hamby et al., 2020). This elucidates the negative influence the spreading of unfair information has.

As suggested by Muhammed and Mathew (2022), the issue of the spreading of unfair information exists in many ITs due to their ability to quickly and broadly spread information. The issue of the spreading of unfair information may be extended to the IT of AI-driven chatbots, as it is known that AI-driven chatbots too can protrude human biases learned in the training data and satisfy confirmation biases thus, they spread unfair information similarly to other ITs (Bashkirova & Krpan, 2024; Caliskan et al., 2017). AI-driven chatbots, also called AI-driven conversational agents (CA), are software programs designed to have, seemingly, natural conversations with the user, which in this study is through a text-based format (Boucher et al., 2021). CAs are built using AI techniques, specifically deep learning neural networks that are used to train the natural language processing (NLP) algorithms of the CA (Lin et al., 2023). Deep learning neural networks allow for processing data in its original form. This is not possible with standard machine learning neural networks, as in standard machine learning it needs to be fed 'features' of the original data (Chauhan & Singh, 2018). Deep learning allows for a much higher detection accuracy making it useful for training NLP algorithms whose goal is to enable computers to interpret human language in a natural way i.e., similar to how humans understand language (Beysolow II, 2018). According to Lin et al. (2023), this allows CAs to grasp and respond to questions better and more naturally.

CAs have risen in popularity in the mainstream population since the release of ChatGPT, this rise has incentivised a lot of researchers to study the applications, merits and pitfalls of CAs (Gordon, 2023; Lin et al., 2023). The capability of CAs to naturally converse makes them a useful tool in, for example, customer service to save costs, or in education as a second teacher (Nicolescu & Tudorache, 2022; Li et al., 2023). Whilst this makes it seem that CAs only bring merit, they also have a flaw. Namely, the aforementioned spread of unfair information may happen in CAs due to their ability to satisfy confirmation biases and protrude human biases (Bashkirova & Krpan, 2024; Caliskan et al., 2017). There are two reasons as to why this may happen. The first is because of a phenomenon called 'hallucinations' (Li, 2023). Due to 'hallucinations', CAs might present information in their answer based on patterns and associations made in the training data under the assumption that the training data is correct. Using such patterns to answer a question results in a loss of context to the question asked. Following this, the answer of the CA will be seemingly

plausible but often partially, or even completely, factually incorrect (Ji et al., 2023; Li, 2023; Siontis et al., 2024). The second reason as to why CAs spread unfair information, according to Li (2023), is due to ‘stochastic parrots’. Here, instead of the CA using the logic and understanding of the training data to deduce an answer, it blindly repeats the data regardless of the fit. This shows the two main factors resulting in a flawed CA’s response.

There are two perceivable ways in which unfair information emerges due to these flaws of CAs. First, it can be through false statements of fact. A common example is how ChatGPT when asked to provide literature tends to create non-existent sources (Siontis et al., 2024). Secondly, as shown earlier, if an NLP model is trained on data that closely reflects human culture including biases or biased information, these biases can protrude in answers of the CA due to the hallucination effect. Thus, creating negative biased assumptions that do not reflect the truth. Often this happens in the form of additional unnecessary judgmental statements that unfairly discriminate against certain groups of people (Caliskan et al. 2017). The danger according to Li (2023) and Zhang et al. (2023) is that the confident answers of the CA which are only partially incorrect or just slightly biased make noticing these flaws difficult which might lead people to take the flawed CA’s answer as true. This enables the conclusion that the result of the hallucination effect combined with the way NLP algorithms for CAs are trained gives rise to the spread of unfair information in a similar way as to the other ITs (Cooke, 2017; Berkman & Katona 2020; Modgil et al., 2021; Muhammed & Mathew, 2022; Valentino et al., 2009). Therefore, from here on out the term unfairness will be used to refer to biased and incorrect information or statements that are elicited in CAs through the hallucination effect.

As recently suggested by Van de Wouw (2023) it is relevant to know how peoples’ attitudes towards CAs are affected through hallucinations which can result in unfairness. Van de Wouw (2023) assessed through a literature review that there has not yet been a study done on the effect of hallucination and thus in a broader sense, unfairness, on the attitude of users towards CAs. This shows the novelty of this line of research. As a result, a preliminary survey was done which concluded that ‘hallucinations’ had little influence on the users’ experience. This could suggest that there is also no influence of unfairness on the user. However, this comes with a caveat, the CA was used in an informal setting where factual relevance was low. Furthermore, research on the influence of unfairness has not yet been done. This creates interest in seeing how user attitudes would be impacted by unfair CAs in formal situations where factual relevance does matter (Van de Wouw, 2023). As there is no conclusive

evidence on the influence of unfair CAs on users, attention is shifted to what attitude factors determine how CAs are perceived by the user.

According to the literature, four main factors are relevant in affecting people's attitudes towards CAs (Følstad & Brandtzaeg, 2020; Zarouli et al. 2018; Zhu et al., 2022). Here, these attitude factors combined form the 'quality of interaction' a person has with a CA, and thus will be named as such moving forward (Shawar & Atwell, 2007). The quality of interaction is a part of the user experience, which refers to the attitudes and perceptions of the users towards the (expected) use which in this context is towards CAs (International Organization for Standardization [ISO], 2019).

The first factor according to Følstad & Brandtzaeg (2020) is usability, which is important for a user's quality of interaction from a pragmatic standpoint. According to the ISO (2018), usability is the achievement of goals in a specific context through a means, in this case, a CA, that is effective, efficient and satisfying according to the user. Usability thus consists of the component of effectiveness i.e., the accuracy and completeness of a user when performing a task with, in this context, a CA (ISO, 2018). A second component of usability is efficiency which is the time and effort needed for the user to be accurate and complete when using, in this context, a CA (ISO, 2018). The final aspect of usability is the satisfaction the user has when, in this context, using a CA. This refers to the extent to which the user does not experience discomfort and has a positive attitude towards the CA. It is the satisfaction dimension of usability that influences the quality of interaction, which it does through three sub-factors (Følstad & Brandtzaeg, 2020; Zarouli et al. 2018; Zhu et al., 2022). Firstly, satisfaction contains perceived usability, which refers to a user's subjective experience regarding the usability of CAs; it improves the quality of interaction when it is high (Følstad & Brandtzaeg, 2020). Secondly, satisfaction contains perceived effectiveness (Zhu et al., 2022). According to Zhu et al. (2022), perceived effectiveness is the confidence a person has for something or someone to perform reliably, in this case, a CA. When the perceived effectiveness is low the quality of interaction also goes down (Kang et al. 2006). Thirdly, satisfaction contains perceived usefulness, which refers to the extent to which a user thinks a CA will lead to better performance (Zarouli et al. 2018). It is known that perceived usefulness has a strong positive relation with the quality of interaction of CAs (Zarouli et al., 2018). When combining all these factors that makeup usability, it shows that usability is positively correlated with the quality of interaction (Mafra et al., 2024).

The second factor, trust, plays a key role in a positive attitude towards CAs and thus a key role in creating a positive quality of interaction (Følstad & Brandtzaeg, 2020). Trust is

generally defined as a relationship between a trustor, a person that is doing the trusting, and a trustee, the CA that must be trusted. This takes place within a specific context. Trust then results from the trustor finding themselves in a situation that causes a lack of control e.g., not knowing information, which is then resolved by the trustee by generating trust through being reliable (Kaplan et al., 2020).

A third factor is perceived competence which, when low, can lead to frustration and be negative to the user's quality of interaction (Følstad & Brandtzaeg, 2020). When perceived competence is high, a person, or in this case a CA, will appear to be professional, responsible, reliable, experienced, credible and intelligent. Lastly, the paper by Zarouli et al. (2018) showed that the fourth factor, perceived helpfulness, has a strong positive relation to the quality of interaction with CAs. Herein, perceived helpfulness is the extent to which a user thinks that the answer provided by the CA can resolve their inquiry (Zarouli et al. 2018). The prior mentioned factors that make up the quality of interaction towards CA's may be influenced based on a CA's unfairness. However, this is not known as research on the influence of unfairness on CA is lacking, making this a novel research subject.

Apart from unfairness, the quality of interaction may also be influenced based on the appearance of the CA; specifically, the sex-based appearance. Følstad et al., (2018) found that trust in the CA is affected by appearance. Female CAs tend to evoke more trust, credibility, perceived uniqueness, morality, positivity and warmth, as compared to their male counterpart. It was found that competence seems to be the same for male and female CAs (Borau et al., 2021). However, these findings on the influence of sex-based appearance differences are debated among researchers. Bastiansen et al. (2022) did not find significant appearance differences in the assessment of competence, trust and helpfulness after people interacted with gendered CAs. They concluded that stereotypical gender biases were not found, which might have been the result of a lack of human likeness in CAs. Human likeness is based on how anthropomorphic the CA is. This pertains to the CA looking, acting, and interacting like a human (Fink, 2012). For human-likeness in CAs, the interaction is the most important; the more a CA communicates like a human in terms of phrasing and minor language mistakes the more human-like it appears (de Sá Siqueira et al., 2023). According to Bastiansen et al. (2022), when CAs become more human-like stereotypical gender biases might also occur. From this, it can be concluded that sex-based appearance differences may influence a user's perception and that it is a relevant aspect when researching the quality of interaction between people and CAs.

Apart from an appearance-based influence, congruence could also influence the quality of interaction. Congruence refers to the sex-based appearance of a CA being equal to the declared sex of the user. It is known that there tends to be a preference for a CA of the same sex as the user (Kim et al., 2021). Jin and Eastin (2023) suggested that it may be because congruence leads to greater satisfaction i.e., usability, specifically in communication. Furthermore, congruence positively impacts the trust in the CA. However, no difference was found in credibility, a part of competence (van Berckel, 2022). Lastly, congruence also tends to have a positive influence on the perception of the information a CA provides, in particular regarding its richness (Sanjeed et al., 2020). These effects of congruence combined suggest that there may be an influence on the quality of interaction.

Following the prior discourse leads to the direction of the paper. As became clear, unfair information spreads fast through various information technologies due to novelty and shock titles, which elicit a lot of clicks and views. Once unfair information is adopted by a person it becomes difficult to change their view (Muhammed & Mathew, 2022; Streicher et al., 2011). The adoption of unfair information has lasting consequences with it often seeing continued use in a person's reasoning whilst they know it is not true, showing the detriment. (de Keersmaecker & Roets, 2017; Hamby et al., 2020). This concept may also take place in human-CA interactions, as CAs, when trained, may adopt human biases that can then be protruded in the CA's answers through the hallucination effect, thus spreading unfairness (Caliskan et al. 2017; Ji et al., 2023; Li, 2023; Siontis et al., 2024). However, there is no conclusive evidence as to whether and how unfairness elicited through CAs has similar consequences for the users or whether it influences users at all; this is a novel research area. It would thus be relevant to see whether people adopt a CA's information as true or can discern it i.e., flag it as unfair. What researchers agreed on is that people tend to base their perceived quality of interaction with a CA on certain key factors e.g., usability, competence, trust, and helpfulness (Følstad & Brandtzaeg, 2020; Zarouli et al. 2018; Zhu et al., 2022). To see if unfairness influences people, looking at the quality of interaction might be a valid starting point. Furthermore, the appearance of CAs may influence a user's perception of the quality of interaction too, even though researchers are divisive on this. In particular, trust may be affected (Bastiansen et al. 2022; Borau et al., 2021). Lastly, congruence between the CA's appearance and the user's declared sex may influence a user's perception of the quality of interaction, albeit debated (Kim et al., 2021; Jin & Eastin, 2023; van Berckel, 2022)

By considering the main factors affecting the quality of interaction with CAs the aim is to explore the potential effect on people's quality of interaction with CAs when these

systems provide a perceivable unfair answer to their questions, have sex-based appearance differences, and are (in)congruent in their sex-based appearance when compared to a person's declared sex. To achieve these goals, the answers to the following research questions are going to be explored and investigated:

RQ 1: Is the ability of people to detect unfairness (i.e., flagging that information is unfair) in the information provided by a CA affected by the CA's appearance (male/female) and congruence between the CA's appearance and participants' declared sex? For this first research question, it may be expected that appearance has an influence. Contemporary research suggests that human-like female CAs may be more trustworthy and credible, thus male CAs might elicit more flagging (Bastiansen et al. 2022; Følstad et al., 2018; Metzger & Flanagin, 2013). Congruence was also found to elicit more trust, thus incongruent CA's may elicit more flagging (van Berckel, 2022).

RQ 2: Is the expected (pre) quality of interaction i.e., measured as the average of the subcomponents of trust, competence, usability, and helpfulness, significantly different from the experienced (post) quality of interaction with a CA, across the manipulated levels of (un)fairness, and as a result of the CA's appearance (male/female), and the congruence between the CA's appearance and the participants' declared sex? For the second research question, it may be expected that across levels of manipulated (un)fairness there are no significant differences in the expected (pre) and experienced (post) quality of interaction. It is known that unfairness may be adopted by the user and is readily believed (Castillo et al., 2011; Muhammed & Mathew, 2022; Li, 2023; Streicher et al., 2011; Zhang et al. 2023). Furthermore, the cause of the spreading of unfair information in CAs, hallucinations, show no influence on the user's quality of interaction, suggesting that unfairness also does not affect the user's quality of interaction (Van de Wouw, 2023). For the appearance it may be expected that human-like female CAs may be more trustworthy than human-like male CAs, thus there may be appearance-based significance regarding the quality of interaction (Bastiansen et al. 2022; Følstad et al., 2018; Metzger & Flanagin, 2013). For congruence, it was found that usability and trust are positively influenced whilst competence showed no difference. Therefore, the quality of interaction overall may be expected to be influenced (Kim et al., 2021; Jin & Eastin, 2023; van Berckel, 2022).

To answer the research questions two studies are conducted. The first study (Study 1) focuses on creating a solid set of stimuli that is clearly distinguishable between fair and unfair which

can be used in this current study but can also be used by others in potential future studies. The need for the first study comes from the novelty of this line of research as conclusive research on unfairness in CAs has not been done before. Because of this, there is no set of unfair stimuli ready and available for use. For creating the stimuli, it is relevant that the negative biased statements containing the unfairness are placed in a context related to the question asked to the CA. This means adding judgements or biased justifications for discriminatory statements that add no relevant information but do not seem out of place in the answer the CA provides (Hardmeier et al., 2021). A general tool that can be used to aid with this is the CRAAP test. With this tool, the currency, relevance, authority, accuracy and purpose of information can be assessed (NJIT, 2021). This enables the creation of well-structured contextually clear answers.

The experimental study (Study 2) will focus on exploring the research questions, through the creation of CAs in the Poe chatbot system with different appearances (male/female) and the ability to answer questions with controlled levels of (un)fairness that are based on the stimuli created in Study 1 (Poe, n.d.).

Study 1 – Stimuli Selection Study

DISCLAIMER: The current study was a collaborative effort of multiple researchers (Lucas Assen, Anna Bader, Nikola Markiewicz, and Seán Verloop). As a result, several components of the thesis, more precisely Study 1 and Study 2 (Design, Participants, Materials, Procedure, and aspects of the Data Analysis) were developed and executed together. Each researcher had access to the same data and collaborated in equal parts on the study design and data collection. The collaboration may have led to textual similarities across the individual works for the design aspects of the study, which is supported by Dr. Simone Borsci.

Methods of Study 1

Participants

A total of 30 participants were recruited, all of whom had given their informed consent before 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. Despite this age range, most of the participants were in their twenties. This can be shown through the median $Mdn = 23.5$ and the interquartile range of IQR [21, 32]. The 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 through online advertising (See Appendix A). For inclusion in the study, participants had to be 18 years or older and be proficient in English. Before the start, the University of Twente's ethics committee assessed and agreed on the execution of the study under request number 240189.

Materials

The goal of the first study was to create and assess the quality of a set of unfair stimuli. The quality was determined by the correctness of the participants' responses, as well as their confidence level in answering. To create testable stimuli, unfair statements were generated on a fictional topic. The choice of using a fictional topic was made to avoid pre-existing knowledge interfering with the quality of the stimuli, as pre-existing knowledge can skew responses (Micocci et al., 2021). The topic used was a Master's programme in Biomechanical engineering at a hypothetical university called ACME, based in the Netherlands. Information on the existing Biomechanical Engineering Master at the University of Twente (UT) was used as a basis (University of Twente [UT], n.d.). Since CAs rely on input from a user to create an output i.e., give unfair information, 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 question, 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. This allows for the quality of the to-be-chosen stimuli's fair and unfair versions to be the same.

To ensure that the unfair responses were indeed unfair, they were created in line with the work of Hardmeier and colleagues (2021), who created a framework of recommendations for preparing unfair problematic information. In the context of CAs, they propose that CAs' pseudo-generated content should contain a regular response to the user in combination with a negative generalisation, a justification of an 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 which serves as a response to the previously presented question was: 'There is no Numerus Fixus for this programme. However, the university aims to take in at least 70% of students from Western European countries, as students from Eastern Europe obviously have a lower work ethic.'.

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

The 32 stimuli were tested using an online survey created in the Qualtrics Survey Software (Qualtrics, n.d.). It included informed consent, demographic questions, an English skill assessment, instructions, and finally the 32 stimuli. Regarding the demographics, participants were asked to state their nationality, age, sex assigned at birth, and gender identity. As for the skill assessment, it was included to ensure that participants were able to understand and accurately respond to the study's materials. They were asked to state their English comprehension skills, their English reading ability, and whether they had any English certificates. The instructions consisted of a scenario and a clarification of the task at hand (See Appendix C). The scenario presented the fictional topic of the Biomechanical Engineering Master's programme at the University of ACME. More specifically, participants were asked to imagine that a random student is considering applying to the Master's. As a result, the student is looking for information about the Master's and asks questions to an AI system on the university's website to gain the information. The scenario was designed to facilitate and simulate a realistic situation wherein potential applicants would seek for such information. The task for the participant was to act as a reviewer of the AI system that would provide a potential applicant with answers to the created questions i.e., the stimuli. As a reviewer, they were asked to flag the pseudo-generated AI answer to be either fair or unfair, and if deemed unfair, to provide a reason why. Furthermore, they were asked, on a five-point Likert scale, to state their confidence in their decision to flag an answer as fair or unfair.

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 responses are

consistently measured across the two stimuli conditions. The 32 stimuli were presented to each participant in a fully randomised order to reduce order effects. For each stimulus, the participants also randomly received either the fair or the unfair version of the CA's answer to the given question.

The gathered participants were provided with the online survey. Upon starting the survey, participants were given the informed consent form which 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 explanation of the task, 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 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 CAs. Finally, their responses were saved, and the survey was completed.

Data Analysis

For both groups of researchers, the study yielded one stream of data through the online survey. The data was exported out of Qualtrics, and into Excel. Here, the data was combined, screened, and filtered. To select the stimuli correctness was used i.e., peoples' ability to correctly flag an answer as fair or unfair. This was done to establish if the stimuli, i.e. answers to the questions, were correctly discriminable. Questions were considered as correctly discriminable when the fair and unfair answers were both correctly categorised by on average more than 95% of the participants. To select stimuli, first, those with the highest correctness were chosen. Then afterwards the average confidence of the participants in flagging a stimulus as fair or unfair combined was used in case some stimuli had equal correctness. The average confidence is the sum confidence of the participants in correctly assessing the stimuli, divided by the total amount of 32 stimuli. The reasoning is that higher average confidence means that the participants had an easier time detecting the unfairness in these stimuli, thus making them more suitable than the others. After the stimuli selection, the stimuli were improved based on the textual feedback of the participants, which was provided if they considered a stimulus unfair.

Results of Study 1

The 32 stimuli were ordered based on the average ability of participants to correctly flag the stimuli as fair and unfair (See Table 1). The best six of the 32 stimuli were selected to

keep the duration of the second study to a minimum and ensure response quality. This was done as the length of the first study was one of the participants' main complaints. The stimuli that were chosen were those that had a level of correctness above 95%. This resulted in stimuli 13, 16, 22, and 25 coming out as the best. Their percentage of correctness, fair and unfair combined, was 100% (See Table 1). The next best stimuli were 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
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	%
S13	15	100	0	0	13	100	0	0	100
S16	16	100	0	0	12	100	0	0	100
S22	15	100	0	0	13	100	0	0	100
S25	13	100	0	0	15	100	0	0	100
S10	16	100	0	0	11	92	1	8	96
S17	14	100	0	0	13	1	93	7	96
S20	14	100	0	0	13	93	1	7	96
S21	15	100	0	0	12	92	1	8	96
S26	14	100	0	0	13	93	1	7	96
S5	12	92	1	8	14	93	1	7	93
S27	13	93	1	7	13	93	1	7	93
S4	15	100	0	0	11	85	2	15	92
S14	15	100	0	0	11	85	2	15	92
S28	11	85	2	15	15	100	0	0	92
S1	12	100	0	0	13	81	3	19	91
S2	12	86	2	14	13	93	1	7	89
S8	13	87	2	13	12	92	1	8	89
S11	11	73	4	27	13	100	0	0	87
S15	13	100	0	0	11	73	4	27	87
S3	13	93	1	7	11	79	3	21	86
S7	12	80	3	20	12	92	1	8	86
S12	13	93	1	7	11	79	3	21	86

Stimulus	Correct fair		Incorrect fair		Correct unfair		Incorrect unfair		Average correct
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	%
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 the stimuli 13, 16, 22, 25 have the highest correctness percentage and that the stimuli 10, 17, 20, 21, and 26 are still within the set range of 95% correctness.

As the goal was to select six stimuli, only two of the five with a correctness of 96% were needed. To determine which two, the participants' average confidence of the fair and unfair answers of the stimuli combined, was used. This refers to the measurement of confidence of the participant in their flagging that was included in the survey after assessing a stimulus. Following this, stimuli 17 and 20 were selected as they had the highest average confidence of the five. The average confidence was 4.54 for both on a 5-point Likert scale (See Table 2). Furthermore, the content of the final six stimuli (13, 16, 17, 20, 22, and 25) showed that unfairness through Nordicism, which was present in four, 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
	<i>n</i>	<i>M</i>	<i>n</i>	<i>M</i>	<i>M</i>	<i>SD</i>
S13	15	4.67	13	4.61	4.64	0.49
S17	14	4.21	14	4.86	4.54	0.69
S20	14	4.50	14	4.58	4.54	0.51
S24	13	4.46	15	4.60	4.53	0.58
S22	15	4.27	13	4.69	4.48	0.74
S25	13	4.08	15	4.80	4.44	0.79
S28	13	4.23	15	4.60	4.42	0.74
S27	14	4.29	14	4.50	4.39	0.63
S30	14	4.00	14	4.79	4.39	0.79
S8	15	4.53	13	4.23	4.38	0.83
S5	13	4.07	15	4.67	4.37	0.74
S31	13	3.92	15	4.80	4.36	0.91
S21	15	4.20	13	4.46	4.33	0.67
S4	15	4.40	13	4.23	4.32	0.67
S10	16	4.43	12	4.17	4.30	0.67
S11	15	4.07	13	4.54	4.30	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	4.20	0.99
S12	14	4.29	14	4.07	4.18	0.90
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	4.10	0.72
S15	13	4.38	15	3.80	4.09	1.12
S3	14	4.14	14	4.00	4.07	0.94
S29	15	3.93	13	4.15	4.04	0.79
S18	15	4.33	13	3.69	4.01	1.10
S14	15	3.67	13	4.08	3.87	0.80

Note. The first column shows the number of participants in the fair condition and their confidence in their assessment. The second column shows the same as the first, but for the unfair condition. The third column shows the average confidence across all participants, for each participant, regardless of condition. This is used to select the remaining two stimuli. 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 suggestions from the participants were used to review and improve the wording and style of the stimuli. The goal was for the stimuli to be more in line with what is expected from a CA i.e., making the answer appear more AI-generated than human-generated. This was done through rephrasing with the help of DeepL and Grammarly. Apart from that, terms present in the stimuli that may not be known to everyone were changed or explained e.g., Numerus Fixus.

Study 2 – Experimental Assessment of Interactions with (Un)fair CAs

Methods of Study 2

Design

Like the first study, the second study was designed and executed by the same group of four researchers, each with their own research questions. This was done to be able to collect a larger set of data. As a result, additional data outside the scope of this paper were collected; only the data relevant to this paper's research questions was analysed.

Following the collaboration, a pre-post (quality of interaction), between-subjects design was employed that was 2 (appearance of the CA: male or female) by 3 (levels of manipulated (un)fairness: fair, 50% unfair, or completely unfair), resulting in six experimental conditions. All independent variables i.e., appearance of the CA, congruence and levels of manipulated (un)fairness, were between-group variables. Due to the pre-post design, the quality of interaction sub-components i.e., trust, usability, helpfulness and competence were measured twice for each participant, once before interacting with the CA, to see their general expectation regarding the quality of interaction with CAs, and once after to see their quality of interaction after the usage. This enabled seeing if the expectation (pre) and experience (post) differ due to levels of manipulated (un)fairness, appearance of the CA and congruence.

Experimental conditions

Six different CAs were designed using the Poe chatbot system i.e., the experimental conditions (Poe, n.d.). This was done by combining the different levels of manipulated (un)fairness and types of appearance. The appearance of the CAs was made discernible using

different pronouns and gendered profile pictures (see Table 3 & Appendix E). In addition to the varying appearance, the CAs were also designed with three different levels of manipulated (un)fairness. The CAs were either completely fair i.e., 0 out of 6 questions unfair, 50% unfair i.e., 3 out of 6 questions unfair, or completely unfair i.e., 6 out of 6 questions unfair (See Table 4).

Table 3.

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

Appearance	Profile	Declaration of gender
Male	Picture of a Caucasian man	“Hi, I am OXX, pronouns He/Him.”
Female	Picture of Caucasian woman	“Hi, I am OXX, pronouns She/Her.”

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

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

Participants

For the second study, 55 participants were recruited; all had given informed consent beforehand. Ultimately, three people were excluded reaching an eventual sample size of 52. One participant was removed due to misunderstanding the instructions of the study. The other two were removed due to not completing the experiment. In the sample, 22 were male and 30 were female, with a mean age of $M=25.15$ years, which ranged between 18 and 50 years. Like the first study, most participants were in their twenties, as the median was $Mdn= 23.50$ with an interquartile range of IQR [21.00, 24.25]. For the experiment, participants were recruited via a non-probability sampling through a mix of voluntary response and convenience sampling, i.e., participants were approached by the researchers on campus or recruited through their acquaintances. Participants had to be 18 years or older and be proficient in English to be included in the study. Furthermore, participation in the first study was an exclusion criterion for the experiment. Before the start of the second study, the University of Twente's ethics committee assessed and agreed on the execution of it under request number 240189.

Materials

The CAs were designed in the Poe chatbot system (Poe, n.d.). The stimuli generated and validated in the first study of the research were inserted into the system and served as the CA's knowledge on how to answer questions. Whether a stimulus's fair or unfair version was used depended on the condition (See Table 4). Similarly, depending on the condition, the CA was primed to present itself with either male or female pronouns and either had a man or woman as a profile picture (See Table 4). Furthermore, the Poe chatbot system could base its answers on additional, general information about the hypothetical Biomedical Engineering Master of the first study (See Appendix E). This was again based on the Biomedical engineering master at the UT (UT, n.d.). The CAs were built around aiding people in finding information about said specific Master as the stimuli, created in study 1, were built around that as well.

A survey was created using the Qualtrics Survey Software (Qualtrics, n.d.). This survey included an introduction, informed consent, and a series of demographic questions regarding nationality, age, sex, and gender identity. It also included the same skill assessment regarding the English language as the first study i.e., asking participants for their comprehension, reading ability and certificates in English. Then, a question assessing whether participants had used CAs before, followed. If answered yes, users were asked to report their frequency of use.

To assess the (expected) quality of interaction with CAs, the individual variables that make up the quality of interaction for both the pre- and post-assessment were measured using Likert scales ranging from 1 to 7 (See Appendix D). The following scales from Bastiansen et al. (2022) were used: a scale measuring the ‘trust’ in the CA consisting of five items (Cronbach’s alpha $\alpha = 0.92$), a scale measuring the ‘helpfulness’ of the CA consisting of five items (Cronbach’s alpha $\alpha = 0.95$), and a scale measuring the ‘competence’ of the CA consisting of five items (Cronbach’s alpha $\alpha = 0.92$) (See Appendix D). Usability, the final sub-component of the quality of interaction, was measured by items 3 through 11 of the Chatbot Usability Scale (BUS-11) (Cronbach’s alpha $\alpha = 0.89$) (Borsci et al., 2022) (See Appendix D). The first two items were removed because they relate to the detectability and findability of a CA, which is irrelevant as access is provided through the survey.

Furthermore, when measuring the expected (pre) quality of interaction, the items of the scales were formulated to refer to the participant’s expectations for a CA based on prior usage of CAs (See Appendix D). If a participant had never used a CA before they were asked to base their responses on their expectations of what a good CA should be; the items of the scales were then formulated accordingly (See Appendix D). In the experienced (post) quality of interaction assessment, the items of the scales were formulated as referring to the CA the participants used in the experiment (See Appendix D). All the items of each individual scale were averaged to obtain a single score for each sub-component of the quality of interaction for both the pre- and post-assessment. To get the ‘quality of interaction’ all individual scales were combined to get two single scores for the quality of interaction, one for the pre- and one for the post-assessment. Lastly to obtain the difference between the expected (pre) and experienced (post) quality of interaction, the delta Δ of the quality of interaction was taken i.e., post minus pre-assessment.

Next, an imaginary scenario was created for the participants to follow during the experiment. In the scenario, the participant acted as a reviewer of a new CA system for the university’s Biomechanical Engineering Master’s webpage. They had to assess the CA by interacting with it and asking it six commonly posed questions by prospective students i.e., the selected stimuli of study 1 (See Appendix C). The main section of the survey provided participants with the six stimuli questions chosen in the first study and a text box to paste the CA’s answer in. This text box was added so it could be confirmed that the CAs behaved as intended. A 5-point Likert scale ranging from ‘strongly disagree’ to ‘strongly agree’ was used to assess whether the provided answer of the CA met the expectations of the participants or whether the participants considered the answer of the CA unfair. 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 or unfairness”) or write their own explanation in a text entry box. The measure of ‘total flagging’ i.e., the detected unfairness by participants, was derived by considering responses from ‘neutral’ to ‘strongly disagree’ as flagged. This resulted in scores ranging from 0 to 6 depending on the number of flagged stimuli by the participant.

Additionally, users were asked to report their intention of use, as measured by a Net Promoter Score (NPS). This was done using the standard NPS question available in the Qualtrics Survey Software and adapting it to the context i.e., CAs: “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?” (Qualtrics, n.d.). To fill in the questionnaire participants required a laptop or a stationary computer with access to the internet.

Procedure

The experiment took place both online and in person, depending on the availability and proximity of the participants. In the online scenario, the researchers connected remotely with the participants to provide them with the login details for the Poe chatbot system (Poe, n.d.). 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. Then the participants were informed about the questions and tasks they would complete, the approximate length of the study, and any potential risks associated with participation. They were warned that the CA may provide problematic output including unfair information. Finally, it was indicated to the participants that their participation was entirely voluntary and that they could withdraw from the study at any time. After listing the contact details of the research team, the participants were asked to confirm that they understood all the information and to give informed consent.

Afterwards, they gave their demographic information, followed by the English skill assessment. Next, participants were asked to report whether they had previous experience using CAs, and if so, the frequency of use in the last 30 days before participating in the study. Afterwards, the pre-assessment took place. The participants reported, through the scales, their expected quality of interaction with a CA.

Then, the participants were given the scenario and were instructed to interact with the CA by asking the provided questions about the imaginary Master’s track. They were instructed to copy and paste the provided question into the CA and then copy and paste the

CA's response into the survey. Once the participants got an answer to one of the questions, their task was to report to what degree the provided answer matched their expectations. In case they identified any issues with the answer such as lack of clarity, misinformation, bias, or other, they were asked to report it.

Next, after interacting with the CA, the post-assessment took place. The participants reported, through the scales, their experienced quality of interaction based on the CA they interacted with. Following this, the participants were asked to fill in a Net Promoter Score (NPS). Finally, after answering all the questions, the participants were provided with a disclaimer of the full purpose of the study before their responses were saved and the survey was completed.

Data Analysis

The second study yielded one stream of data that was exported from the Qualtrics Survey Software as a Microsoft® Excel file (Version 16.86) with numeric choice text. Before analysing the data, it was cleaned by removing three participants due to incompleteness and lack of understanding of the instructions. Furthermore, the data was transformed; the scales and scores were aggregated to create the variables necessary for answering the research questions. Next, the cleaned data was imported into the statistical software R Studio® Version: 2024.04.0+735 for R version 4.4.0. For all performed analyses the following packages were used: readxl, tidyverse, psych, car, dyplr, tidyr, ggplot2, janitor, regclass dgof, ggpubr, and performance (See Appendix F).

Firstly, an outlier analysis was performed by calculating Cook's distance within a generalised linear model (GLM). In the GLM, the 'total flagging' was the dependent variable and the levels of manipulated (un)fairness were the independent variable. Potentially finding and removing influential outliers in this model ensures that the main manipulated variable i.e., the levels of manipulated (un)fairness are not disturbed and clearly distinguishable. Therefore, a conservative threshold of $4/n$, whereby n is the sample size, was taken (Blatná, 2006). Afterwards, a first impression of the data could be had by looking at descriptive statistics. This was done by looking at the distribution of the general data through medians, ranges, and outliers of boxplots of the 'quality of interaction' in the post-assessment divided by the experimental conditions i.e., the appearance of the CA and the levels of fairness. Another set of boxplots was created for the delta Δ quality of interaction divided by the experimental conditions i.e., the appearance of the CA and the levels of fairness. The boxplots focussing on the delta Δ quality of interaction represent the distribution of the data regarding the difference between the expected (pre) and experienced (post) quality of interaction, which was

calculated through post minus pre. Boxplots were not created for the quality of interaction in the pre-assessment. The informational relevance regarding differences due to the appearance of the CA and the levels of (un)fairness is non-existent in the pre-assessment as the participants had yet to interact with the CAs. Next to boxplots, a table was created to elucidate the central tendency and variability of the data through the means and standard deviations of the pre- and post-subcomponents of the quality of interaction i.e., trust, helpfulness, competence, and usability. This was again divided by the experimental conditions i.e., manipulated levels of (un)fairness and CA appearance. Lastly, apart from the experimental conditions, the research questions also focus on the influence of the (in)congruence between the CA's appearance and the participant's declared sex. Thus, an additional box plot was created to show the medians, ranges, and outliers for the distribution of the delta Δ quality of interaction over the independent variable of congruence. Furthermore, means and standard deviations were reported for the central tendency and variability.

After getting an impression of the data, the scales used to measure the quality of interaction variables were checked for reliability using Cronbach's alpha α . The reliability was individually checked for the pre- and post-assessment as it is considered good practice to check it for each use of a scale. Based on a literature review by Taber (2018) the commonly accepted and adopted acceptable threshold for Cronbach's alpha is $\alpha > .70$. For that reason, it has also been set as the threshold in this study. In case of alpha lower than $\alpha < .70$ it will be considered on a case-by-case basis as to how to proceed. In general, an alpha between $\alpha = .50-.59$ is considered poor and an alpha between $\alpha = .60-.69$ is considered questionable (Taber, 2018). The Cronbach alphas were also compared to the original literature in which the scales were created.

Following the assessment of the reliability of the scales, it had to be assessed whether the parametric assumptions were met. The first parametric assumption requires the dependent variables in the data to be continuous. Whilst most variables in this study were measured through Likert scales and thus would be considered ordinal, this should not pose an issue as the consensus is that Likert scales with 5 points or more, which is the case, can be used for parametric testing. The ordinal scales then approximate a continuous variable (Johnson & Creech, 1983; Sullivan & Artino, 2013). The second parametric assumption pertains to the normality of the data i.e., whether the data is normally distributed around the mean. This was tested using a Shapiro-Wilkes test which tends to be the best measure for testing non-normality according to Shapiro et al. (1968). The data was considered normal when the test showed non-significance, thus for an alpha of $\alpha=.05$ the p-value had to be greater than $p>.05$.

The normality was tested for all numerical variables used in the exploratory analyses. The third parametric assumption is the homogeneity of variances (homoscedasticity), which assumes that there is an equal amount of variance across the variables. This was checked using Bartlett's test, which is useful when working with linear regression. The test was run over the GLMs in the exploratory analyses to see if the models suffered from non-homogeneity of variances (heteroscedasticity). Bartlett's test assumes equal variance when, for an alpha of $\alpha=.05$, the p-value was greater than $p>.05$ (Arsham & Lovric, 2011). Lastly, the fourth parametric assumption is multicollinearity. Here it is assumed that the independent variables in a model do not correlate, if they do, they are multicollinear. To investigate this the variance inflation factor (VIF) was calculated in the GLMs with multiple independent variables to see if multicollinearity was present. Johnston et al. (2018) advocate for a VIF score below 10 and preferably as low as possible, this advice was followed in this study. Generally, a VIF score of 10 and higher is severely multicollinear, between 5 and 10 is moderately multicollinear, and between 2.5 and 5 is mildly colinear (Johnston et al., 2018).

After checking the parametric assumptions, a manipulation check was performed to ensure that the perceivable levels of (un)fairness were correctly manipulated i.e., to assess whether people were able to discriminate between the different levels of manipulated (un)fairness during the experiment. This was done using a GLM to test whether the levels of manipulated unfairness as the independent factor variable, with completely fair as the reference category, predict 'total flagging' as a dependent variable. Here it is expected that higher levels of manipulated unfairness elicit more flagging. This GLM and the following GLMs were tested for significance based on the z -value or t -value of the individual model parameters which were considered significant for an alpha of $\alpha=.05$ when the p-value was below $p<.05$.

Next, the focus was put on investigating the exploratory research questions. The first research question explores if people's ability to detect unfairness i.e., flagging information as unfair, is influenced by the CA's appearance and congruence with the participants' declared sex. This was tested through a GLM, whereby total flagging was the dependent variable whilst CA appearance and congruence were the independent factor variables. In the GLM, the male version of the CA's appearance and incongruence were the reference categories.

The exploration of the second research question focussed on how the difference between the expected (Pre) and experienced (Post) quality of interaction is influenced by the different levels of manipulated (un)fairness, the appearance of the CAs and the congruence. To test this a GLM was used with the delta Δ quality of interaction as the dependent variable

and the levels of manipulated (un)fairness, the CA's appearance and congruence as independent factor variables. Here complete fairness, the male version of the CA's appearance and incongruence were the reference categories. In the case that any parametric assumptions were not met, the GLMs used the type of regression that applies best to the specific situation which will be mentioned in the results.

Lastly, paired t-tests were done on the quality of interaction and its subcomponents. In the case of non-normal data, a paired Wilcoxon signed-rank test was done. These tests ensure that the differences between levels of manipulated (un)fairness came from a significant difference between the pre and post, within levels of manipulated (un)fairness. It is also an additional check on the subcomponents of the quality of interaction i.e., trust, competence, helpfulness, and usability between their pre- and post. The tests were considered significantly different for an alpha of $\alpha=.05$ when the p-value was below $p<.05$.

Results of Study 2

First, the outlier analysis was performed, which resulted in the removal of 4 participants. Three were in the completely unfair condition and one was in the completely fair condition (See Appendix G). This brought the total amount of participants included in further analyses to 48. Here, 16 were in the completely fair, 18 in the 50% unfair, and 11 in the completely unfair condition. This showed that the participant count was relatively low in the completely unfair condition due to the removal of participants.

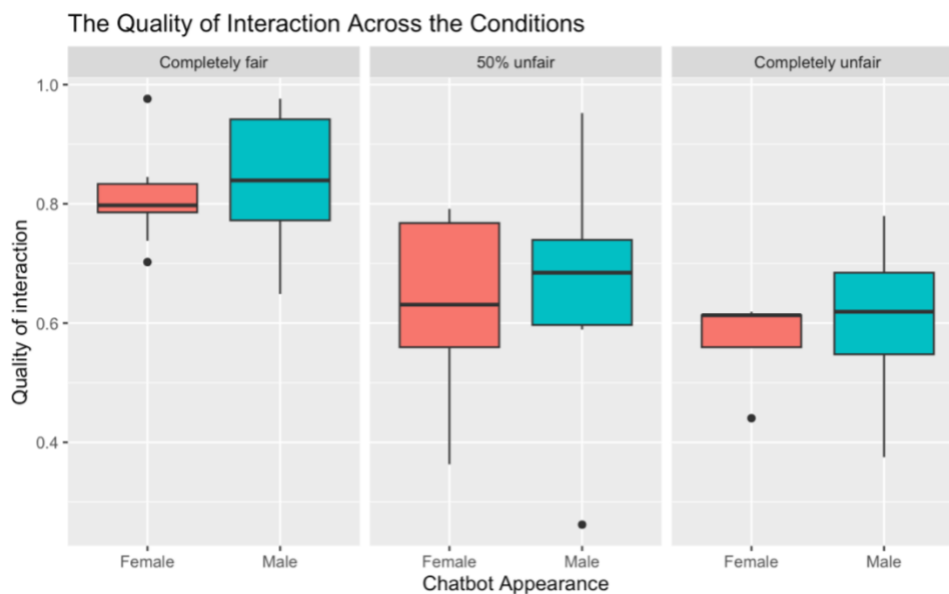
Descriptive statistics for the main variables of the data

Afterwards, the focus was on gaining a first impression of the data. To do so, two boxplots were created. The first box plot showed the medians, ranges, and outliers of the (experienced) quality of interaction in the post-assessment of the participants (See Figure 1). The boxes were separated by the experimental conditions i.e., the manipulated levels of (un)fairness, as well as the appearance of the CA (male or female). The boxplot elucidated that the completely fair condition scored higher on the experienced (post) quality of interaction than the 50% unfair and completely unfair conditions (See Figure 1). Furthermore, for the completely fair condition, it seemed that the scores for the female CA were more centred than for the male CA. This was confirmed by looking at the female CA's median $Mdn= 0.80$ and interquartile range IQR [0.79, 0.83], which showed denser distributed scores when compared to the male version's median $Mdn= 0.84$ with an interquartile range of IQR [0.77, 0.94]. The scores for the quality of interaction in the post-assessment between the 50% unfair and completely unfair condition showed a slight decrease in completely unfair condition (See Figure 1). In the 50% unfair condition, there was a broader spread in the scores

of the experienced (post) quality of interaction for the female CA. This was proven by the median and interquartile range of the female CA, which are $Mdn = 0.63$ and IQR [0.56, 0.77] compared to $Mdn = 0.68$ and IQR [0.60, 0.74] for the male version. For the completely unfair condition, the scores on the total experience post were more densely spread for female CA's ($Mdn = 0.61$, IQR [0.56, 0.61]) compared to the male CA's ($Mdn = 0.62$, IQR [0.55, 0.68]).

Figure 1

Box Plots of Participants' Experienced (Post) Quality of Interaction Divided by the Experimental Conditions i.e., the CA's Appearance and Levels of Manipulated (Un)fairness

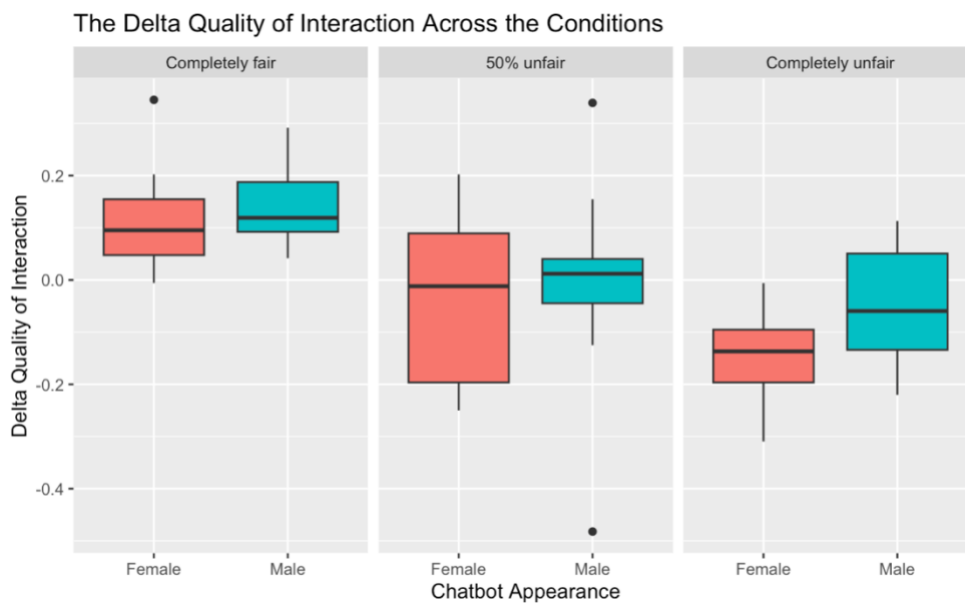


The second box plot showed the medians, ranges, and outliers of the Δ quality of interaction of the participants i.e., the difference between expected (pre) and experienced (post) quality of interaction (See Figure 2). The boxes were separated by the experimental conditions i.e., the manipulated levels of (un)fairness, as well as the appearance of the CAs (male or female). The main observation that could be made was that the completely fair condition had higher positive values than the values of the 50% fair condition, which were distributed around 0. The values for the completely unfair condition were distributed slightly below 0, thus even lower (See Figure 2). Based on the Δ quality of interaction, which is post minus pre, the positive values seen in the completely fair condition showed that the experienced (post) quality of interaction was higher than the expected (pre). The negative values of the completely unfair condition mean the opposite i.e., pre was greater than post. Furthermore, in the completely fair condition, the female CA ($Mdn = 0.10$, IQR [0.05, 0.15]) showed lower Δ scores when compared to the male version ($Mdn = 0.12$, IQR [0.09, 0.19]). In the boxplot, it could also be seen that there was a difference between the 50% unfair and completely unfair conditions, with the completely unfair condition being slightly lower.

The female CA of the 50% unfair condition showed a median of $Mdn = -0.01$, and the median for the male version is $Mdn = 0.01$ meaning that both had their scores spread around 0. This alluded to no differences between the pre and post, for the 50% unfair condition. Furthermore, for the 50% unfair condition, the scores of the female CA were broader spread (IQR [-0.20, 0.09]) when compared to the male version (IQR [-0.04, 0.04]). In the completely unfair condition, the spread of scores was primarily below 0, suggesting lower post than pre scores. The female CA's distribution was also slightly lower ($Mdn = -0.14$, IQR [-0.20, -0.10]) than the male version ($Mdn = -0.06$, IQR [-0.13, 0.05]). This suggested a greater expected (pre) than experienced (post) quality of interaction for the female CA.

Figure 2

Difference (Delta) in Expected (Pre) and Experienced (Post) Quality of Interaction Divided by the Experimental Conditions i.e., the CA's Appearance and Levels of Manipulated (Un)fairness



For the central tendency and variability of the data, the means and standard deviations of the subcomponents of the quality of interaction i.e., trust, competence, helpfulness and usability were reported for both the pre- and post-assessment. Like the boxplots, this was divided over the experimental conditions i.e., the appearance of the CA and the levels of manipulated (un)fairness (See Table 5). Firstly, it was found that the means were all above $M > 0.60$, except for post trust of the female CA in the 50% fair condition, the post trust for both the male and female CA in the completely unfair condition, and post competence for the female CA in the completely unfair condition. Thus generally, regardless of the condition participants rated the CAs on the higher end of the scale. The next finding is that the

completely fair condition for both the male and female CA saw a substantial increase between pre and post. This was in line with the second boxplot which, as mentioned, showed that the Δ quality of interaction was higher and positive in the completely fair condition, meaning that the experienced (post) compared to the expected (pre) quality of interaction is higher (See Figure 2). Such differences could not be seen in the 50% unfair condition, as the mean scores stayed relatively the same or there was a slight decrease (See Table 5). The only exception to this is pre- and post-helpfulness in the 50% unfair condition, which sees an increase of 0.01 for both the male and female CA (See Table 5). In the completely unfair condition, there is a clearer decrease between pre and post. In particular for the difference between pre- and post-competence of the female CA which sees a drop of 0.22.

Table 5

Mean and Standard Deviations for Expected (Pre) and Experienced (Post) Scores of the Quality of Interaction Variables Divided by the Experimental Conditions i.e., the CA's Appearance and Levels of Manipulated (Un)fairness

Variables	Completely fair		50% Unfair		Completely 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.

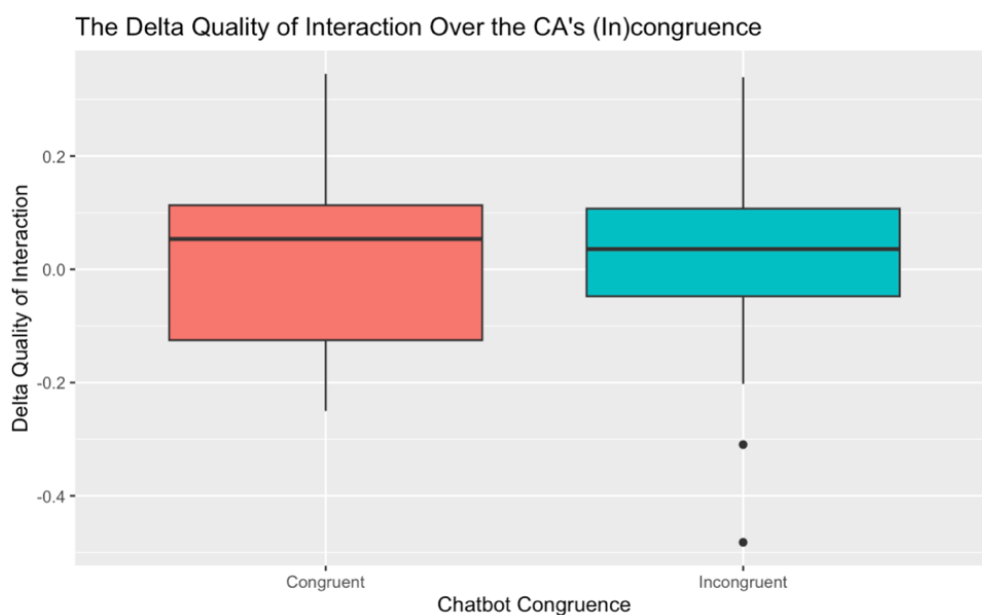
The final box plot focussed on the distribution, central tendency and variability of the Δ quality of interaction over the (in)congruence of the CA's appearance and the

participant's sex (See Figure 3). Overall, the scores were similarly distributed for both congruence and incongruence (See Figure 3). However, incongruence showed a denser distribution around the median, which was reflected in the interquartile range ($Mdn= 0.04$, IQR [-0.05, 0.11]), when compared to congruence ($Mdn= 0.05$, IQR [-0.13, 0.11]).

Furthermore, the central tendency and variability between congruence and incongruence were also similar. This was reflected in the means and standard deviations with them being $M=0.02$ and $SD=0.16$ for congruence and $M=0.01$ and $SD=0.18$ for incongruence.

Figure 3

Difference (Delta) in Expected (Pre) and Experienced (Post) Quality of Interaction Over the (In)congruence of the CA's Appearance and the Participant's Declared Sex



Reliability analysis of the quality of interaction subcomponents

After gaining a first impression of the data, the scales measuring the subcomponents of the quality of interaction i.e., usability, trust, competence and helpfulness, were tested for reliability to ensure that the data was consistent and could be used for further analyses.

Cronbach's alpha was applied to measure the reliability of both the pre- and post-assessments.

Firstly, the BUS scale, which measured usability, consisted of 9 items (Pre; $\alpha = .67$, 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; $\alpha = .91$). Following these results, all scales,

except for the BUS scale in the pre-assessment, passed the threshold of an acceptable level of

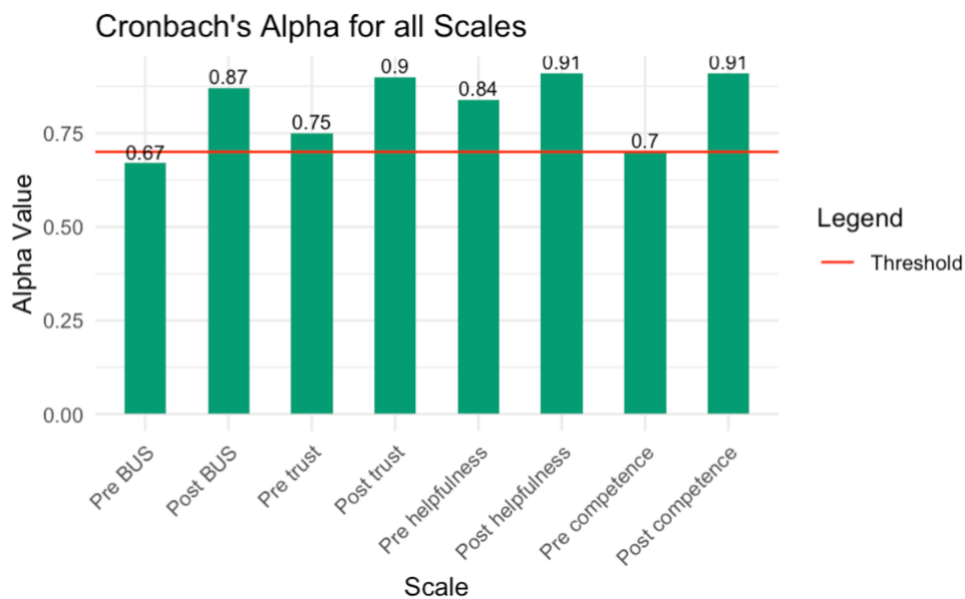
reliability which was set at an alpha of $\alpha > .70$ (See Figure 4). Removing items in the pre-BUS

scale could not improve the value of $\alpha = .67$ to bring its alpha over the acceptable threshold.

However, since it is close to the threshold and thus at the higher end of the range of a questionable alpha, and because the post BUS scale is very reliable it was decided to accept it. An additional remark on the scales is that none reached the level of reliability as measured in their original papers, where the BUS had 11 items ($\alpha = .89$) (Borsci et al., 2022). The original scales of trustworthiness, helpfulness, and competence were 5 items respectively ($\alpha = .92$, $\alpha = .95$, $\alpha = .92$) (Bastiansen et al., 2022). It can then be seen that the Post competence scale is the closest to the original Cronbach's alpha with a difference of .01, whilst the Pre competence scale is the farthest away with a difference of .19. Furthermore, it should be noted that there is quite a gap in the reliability in the pre vs post scales (See Figure 4). The biggest gap, of .21, can be found between the pre- and post-scales of competence, whilst the smallest gap, of .07, can be found between pre- and post-helpfulness.

Figure 4

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



After the reliability of the scales, the parametric assumptions of the data were tested. A Shapiro-Wilkes test was performed to test the normality of the used variables. To test homogeneity of variance (homoscedasticity) and multicollinearity in the GLMs, a Bartlett test and VIF scores were used. All parametric assumptions were met except for the normality of total competence pre and post, total helpfulness pre and post, and total flagging. Furthermore, the manipulation check model was close to violating the homogeneity of variance (homoscedasticity) but did not (See Appendix G). The inferential analyses were done using GLMs, which can handle the non-normal data when accounted for. To do so, for total

flagging specifically, Poisson regression was used. As a result, the non-normality was not considered an issue.

Manipulation check for the levels of (un)fairness

Before exploring the research questions, a manipulation check was performed to ensure that people could discriminate between the perceivable levels of (un)fairness to confirm that the levels were correctly manipulated. The estimate of the completely fair condition, i.e., the intercept, was positive and significant, thus greater than 0. This is unexpected as the completely fair condition did not have items to flag (See Table 6). However, there were significantly higher amounts of flagging for higher levels of unfairness when compared to the completely fair condition i.e., the intercept, thus the manipulation worked (See Table 6). Flagging between the unfair levels i.e., 50% unfair and completely unfair, was also significantly different. This was shown by making the completely unfair condition the reference category. The completely fair condition showed a significantly lower estimate ($B=-3.50$, $z=-5.71$, $p<.001$), and the 50% unfair condition as well ($B=-1.82$, $z=-2.69$, $p=.007$). Thus, the completely unfair condition showed significantly more flagging than both the completely fair and 50% unfair condition, this reaffirms that the manipulation check worked.

Table 6

Overview of the Generalised Linear Model Used for the Manipulation Check, With the Estimates, Standard Errors, z-values, and p-values for Each of the Model's Predictor Variables

Predictors	<i>B</i>	<i>SE</i>	<i>z</i>	<i>p</i>
Manipulation check model				
Intercept (completely 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 the participants flagged as unfair, ranging between 0 and 6.

Exploration and investigation of the research questions

After the manipulation check, the research questions could be investigated. The first research question explores if people's ability to detect unfairness i.e., flagging a question as unfair, is influenced by the CA's appearance and congruence with the participants' declared

sex (See Table 7; Total flagging model). The GLM showed no significant difference due to the CA's appearance, nor were there significant differences due to the (in)congruence of the CA's appearance with the participants' declared sex (See Table 7; Total flagging model). Thus, neither the CA's appearance nor the congruence had a significant influence on the participants' ability to detect unfairness i.e., flagging information as unfair.

Through the second research question, it was investigated whether the difference in expected (pre) and experienced (post) quality of interaction i.e., the delta Δ quality of interaction, is significantly different across the manipulated levels of (un)fairness and influenced by the CA's appearance and the congruence. To establish this a GLM was created that looked at the differences in pre and post using the delta Δ quality of interaction (See Table 7; Delta Δ quality of interaction model). The model showed that the intercept is significant and positive, thus the reference category i.e., the completely fair condition, is greater than 0 (See Table 7; Delta Δ quality of interaction model). Considering the meaning of the delta Δ quality of interaction variable this suggests that the experienced (Post) quality of interaction was greater than the expected (Pre) quality of interaction. Furthermore, the 50% unfair and completely unfair conditions showed significant negative estimates. This meant that they had significantly lower delta Δ quality of interaction scores compared to the completely fair condition. Thus, there is significant evidence that the delta Δ quality of interaction differs across levels of manipulated (un)fairness (See Table 7; Delta Δ quality of interaction model). When changing the reference category from the completely fair condition to the completely unfair condition, it could be seen that complete fairness still showed a significant difference compared to complete unfairness ($B=0.22$ $t= 4.07$, $p<.001$). However, a significant difference was not found in the 50% unfair condition when complete unfairness was the reference category ($B=0.07$, $t=1.32$, $p=.193$). Thus, whilst there was a significant difference in the delta Δ quality of interaction across the completely fair and unfair conditions, across the 50% unfair and completely unfair condition there was no significant difference.

Apart from fairness, the CA's appearance was also considered as a potential influence factor on the delta Δ quality of interaction; however, no evidence was found that there was a significant difference in the delta Δ quality of interaction between the male and female versions (See Table 7; Delta Δ quality of interaction model). Similarly, it was postulated that the congruence between the CA's appearance and the participant's declared sex influenced the delta Δ quality of interaction; however, no evidence was found that there was a significant difference (See Table 7; Delta Δ quality of interaction model).

Table 7

Overview of the Generalised Linear Models With Their Estimates, Standard Errors, t-values or z-values, and p-values for Each of the Models' Predictor Variables

Predictors	<i>B</i>	<i>SE</i>	<i>z</i>	<i>p</i>
Total flagging model				
Intercept (male, incongruent)	2.35	0.36	6.53	<.001***
Appearance (female)	-0.02	0.42	-0.04	.970
CA congruence (congruent)	-0.43	0.43	-0.99	.321
Predictors	<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Delta Δ quality of interaction model				
Intercept (completely fair, male, incongruent)	0.16	0.05	3.32	.002**
50% unfair	-0.15	0.05	-3.14	.003**
Unfair	-0.22	0.06	-4.07	<.001***
Appearance (female)	-0.05	0.04	-1.09	.284
CA congruence (congruent)	-0.003	0.04	-0.07	.942

Note. The total flagging model is using Poisson regression to deal with the non-normality of the 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.

The last performed analyses, dependant on whether the data was parametric or not, were either paired t-tests or paired Wilcoxon signed-rank tests between the pre- and post-conditions for the quality of interaction and its subcomponents i.e., trust, competence, helpfulness, and usability. Based on the GLMs it came to light that there were significant differences in the delta Δ quality of interaction across the different levels of manipulated (un)fairness. The paired t-tests and paired Wilcoxon signed-rank tests clarified the reason for the significance. They showed that there were significant differences between the expected (pre) and experienced (post) quality of interaction and its sub-components for the completely fair condition and for some of the completely unfair condition (See Table 8). Contrarily, none of the 50% unfair pre- and post-assessment paired t-tests and paired Wilcoxon signed-rank tests showed significance (See Table 8). Furthermore, these significant effects combined with the descriptive statistics showed that the completely fair condition saw significant increases in

the quality of interaction and its subcomponents from pre to post, whilst the completely unfair condition saw slight significant decreases (See Table 5). These findings revealed and reconfirmed that there are indeed significant differences in the quality of interaction between pre and post and across the levels of manipulated (un)fairness. Namely, between the completely fair and the unfair conditions. There is no difference between the 50% unfair and completely unfair conditions, which is surprising as the quality of interaction in the completely unfair condition significantly differs between pre and post. Furthermore, the findings showed that these significant differences across levels result from an increase in the quality of interaction due to complete fairness, and a decrease due to complete unfairness.

Table 8

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

Variables	Completely fair		50% Unfair		Completely unfair	
	<i>t(df)</i>	<i>p</i>	<i>t(df)</i>	<i>p</i>	<i>t(df)</i>	<i>p</i>
Pre vs Post quality of interaction	-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*
	<i>V</i>	<i>p</i>	<i>V</i>	<i>p</i>	<i>V</i>	<i>p</i>
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 with the degrees of freedom between brackets and the p-values. This has been done for the pairwise t-tests and pairwise

Wilcoxon signed-rank tests between the pre- and post-assessments of the quality of interaction and its subcomponents over the different levels of fairness.

Discussion

This study started by delineating how unfair information spreads through ITs, and what its consequences are. For example, the satisfaction of confirmation biases leading to a person falling into a filter bubble (Modgil et al., 2021; Valentino et al., 2009). A link was made with CAs as they, as an IT, also spread unfair information through the hallucination effect (Caliskan et al. 2017; Ji et al., 2023; Li, 2023; Siontis et al., 2024). However, the study of unfairness in CAs is novel and the influence of unfair CAs on people and what factors play a role in this is unknown. Therefore, to make a start in this novel line of research, the goal was to see how a person's quality of interaction with a CA is influenced. This was done by looking at the influence of (un)fairness. Furthermore, it may also be that the CA's appearance and the congruence of the CA with a participant's declared sex influence the quality of interaction, thus these were included as well (Bastiansen et al. 2022; Følstad et al. 2018; Jin & Eastin, 2023; van Berckel, 2022). This led to two research questions. The first explores whether the CA's appearance and congruence influence the ability to detect unfairness, i.e., whether people start flagging differently. The first research question shows that neither the CA's appearance nor the congruence has a significant effect on the participants' ability to detect unfairness i.e., their amount of flagging. The second research question pertained to whether the difference in expected (pre) and experienced (post) quality of interaction with a CA is significantly different across levels of manipulated (un)fairness. Furthermore, it looked at whether the difference is influenced by the CA's appearance and congruence. The findings show a significant difference in the expected (pre) and experienced (post) quality of interaction across levels. Namely, between the completely fair and both unfair conditions, but not between the 50% unfair and completely unfair conditions. The findings also showed no differences between the expected (pre) and experienced (post) quality of interaction due to the difference in the CA's appearance or congruence. After summarising the main findings of the research questions, an in-depth look is taken to uncover the takeaways of the findings considering prior literature.

The first research question explored whether the appearance and congruence of the CA influence the ability to detect manipulated (un)fairness i.e., whether people flag more or less. In the literature, the existence of an effect of appearance is debated with Borau et al. (2021) and Følstad et al. (2018) stating there is, whilst Bastiansen et al. (2022) found no such effect.

Based on Borau et al. (2021) and Følstad et al. (2018) it was expected that male CA would elicit more flagging due to a potentially lower trustworthiness and credibility. However, the results show that this expectation is not met, there was no significant difference in flagging, regardless of whether the CA was male or female. Thus, the findings are in line with Bastiansen et al. (2022) in this debate. Namely, in this study, no sex-based appearance differences were found. The level of (un)fairness, according to users, is thus expected to be the same regardless of appearance.

Similarly to the appearance of CAs, congruence was expected to show less flagging due to higher trust compared to incongruence (van Berckel, 2022). However, van Berckel (2022) also pointed out that for the factor credibility, no differences exist between congruence and incongruence, which could suggest no differences in the ability to detect fairness through congruence. This is also what is supported based on the findings of this study, there are no differences due to congruence. The findings thus support that the alignment between a CA's appearance and a user's declared sex does not influence users to flag differently as compared to when there is no alignment; the level of (un)fairness, according to the users, is thus expected to be the same.

Following the exploration of the first research question the findings of the second research question can be investigated. Firstly, the focus is on the main novelty of the study i.e., whether the difference in expected (Pre) and experienced (Post) quality of interaction with the CA is significantly different across levels of manipulated (un)fairness. It was expected that there would be no difference as it is known that the user may adopt unfairness and readily believe it (Castillo et al., 2011; Muhammed & Mathew, 2022; Li, 2023; Streicher et al., 2011; Zhang et al. 2023). Furthermore, the cause of the spreading of unfair information in CAs, hallucinations, shows no influence on the user's quality of interaction, suggesting that unfairness also does not affect the user's quality of interaction (Van de Wouw, 2023). The results were partially in line with this whilst also being somewhat unexpected.

In line with the expectations is the finding that the differences between expected (pre) and experienced (post) quality of interaction are not significantly different across the unfair conditions. However, against expectations, there is a significant difference between complete fairness and the two unfair conditions. This effect mainly came from an increase in the experienced (post) quality of interaction within the completely fair condition. The reason is that an increase or decrease within a level of (un)fairness changes the delta Δ quality of interaction, which then affects the difference across levels. A second unexpected result is that within the completely unfair condition, there is a slight significant decrease in the quality of

interaction between pre and post. However, as mentioned earlier this slight decrease in pre and post in the completely unfair condition was not large enough for it to be significantly different from the pre- and post-difference in the 50% unfair condition.

Combining the results of the novelty aspect of this study, it shows that across levels of manipulated (un)fairness differences in the Δ quality of interaction only occur between the completely fair and both unfair conditions, not across all levels. Furthermore, differences within the levels are due to an increase in the quality of interaction in the completely fair condition and a slight decrease in the completely unfair condition; no differences were found in the 50% unfair condition. This is partly in line with the literature that there would be no differences due to unfair information (Castillo et al., 2011; Muhammed & Mathew, 2022; Li, 2023; Streicher et al., 2011; Van de Wouw, 2023; Zhang et al. 2023). The main unexpected finding is that when a CA is completely fair it results in a higher quality of interaction than was expected by the user, whilst a completely unfair CA results in a slightly lower quality of interaction than was expected by the user. These results show the practical relevance of this novel study; it provides evidence that accounting for fairness when designing a CA is important in order to deliver a high 'quality of interaction' for the user.

To deepen the understanding, the results imply that people *expect* CAs to be unfair to some extent. As a result, the quality of interaction is rated no different in the pre-and post-assessment for the 50% unfair condition and only slightly lower in the post-assessment for the completely unfair condition. This is in line with ample literature that shows that, especially in young adults, people are aware of, consume, and spread unfair information through social media. Therefore, showing that these groups may expect the presence of unfairness (Balakrishnan, 2022; Hargittai et al., 2010). The results suggest that this expectation may also exist in CAs. Since the quality of interaction increased in the completely fair condition, it may be that people are pleasantly surprised by the absence of unfairness, thus showing a greater experience than expected.

The second research question also postulated that there is an influence of the appearance of CA on the Δ quality of interaction. Again, whilst the effects of appearance are debated it was expected that human-like female CA's may be more trustworthy than human-like male CA's, thus the expectation was that there may be an appearance-based difference (Borau et al. 2021; Følstad et al., 2018; Metzger & Flanagin, 2013). However, no difference in the Δ quality of interaction was found due to the appearance of CAs. These findings are again in line with Bastiansen et al. (2022) who also found no appearance-based

differences in CAs. Thus, it can be concluded that no differences between the expected (pre) and experienced (post) quality of interaction are the result of the CA's appearance.

Lastly, it was postulated that congruence between a CA's appearance and a user's declared sex could affect the Δ quality of interaction. Congruence was expected to show a higher Δ quality of interaction according to Kim et al. (2021) and Jin and Eastin (2023). However, van Berckel (2022) suggested that there may be no differences at all, based on some aspects that make up the quality of interaction. The current study is in line with van Berckel's (2022) findings; no differences were found. This is likely in part because there are also no appearance-based differences. Thus, it shows that any differences between the expected (pre) and experienced (post) quality of interaction are not due to the CA's (in)congruence with the participants' declared sex.

The Novelty of Study 1

Apart from study 2 answering the research questions, there was also study 1. This study was done due to the novelty of this line of research; no usable set of validated stimuli was available, thus highlighting the need for the first study. This study focused on the creation and validation of a set of robust fair and unfair stimuli that were to be used in the second study and could be used in future studies. This resulted in six 'best performing' stimuli, all with a fair and unfair version. Within the six stimuli, two had a bias regarding sexism and four had a bias regarding Nordicism i.e., discrimination based on a supposed superiority of north-western Europeans. It is in line with the literature for these stimuli to come out as best performers. According to Tomaskovic-Devey and Mccann (n.d.) unfair treatment is most often perceived and reported through race-based discrimination with sex-based discrimination being a close second. It follows logically that these stimuli were better performers than those that included other biases, e.g., ageism.

Limitations of the Study and Recommendations for Future Works

Following the novelty of study 1 limitations and solutions of the studies are discussed. The first limitation took place in the stimuli selection study. The biases were one-way, i.e., stimuli containing sexism did so against women and stimuli containing Nordicism did so against south-eastern Europeans; this does not prove that Nordicism and sexism would be best performers if the direction was reversed. It follows that a recommendation for future studies is to see if the reversal of the bias direction still yields the same results.

The second limiting factor was the removal of four participants due to the outlier analysis. Three were in the unfair condition seriously reducing the number of participants in this condition. The reason was that in the unfair condition, the outliers did not flag anything

(See Appendix G). This behaviour is odd and likely due to a design limitation of the study. It likely has to do with the response generation of the Poe chatbot system (Poe, n.d.). In the completely unfair conditions, on some occasions, though unclear why, the CAs made the user aware to be cautious of incorrect information, although it did not state that it was referring to its own answers. This may have made people consider the CA as truthful and fair as they, seemingly honestly, were trying to warn users of the potential dangers of using a CA. In future studies this could be prevented by using an alternative to the Poe chatbot system, ideally self-build. Furthermore, results in the unfair condition should be accepted with caution as the removal made the sample size of the unfair condition small. A future study should reaffirm the effects with a larger sample.

A third limitation was the difference in reliability between the pre- and post-scales. Especially for the BUS scale, for which the pre-assessment was considered questionable. There can be both external and internal factors that contribute to such a reliability difference between pre and post (Ursachi et al., 2015). In the post-assessment, participants answer the scales with the experience they just had in mind. In the pre-assessment, the participants are asked to assess their expected experience with a CA based on prior experiences with other CAs. If the participants had never interacted with a CA before, they were asked what they would expect of a CA in general. When the participants answer the scales with multiple CAs in mind they might answer differently on the individual items depending on the CA they think of. The scales may then measure the quality of interaction of multiple CAs i.e., objects, at the same time which may lower the internal consistency and thus reliability (Ursachi et al., 2015). Although it did not have a serious impact, in future studies, participants should be asked to base their answers on the pre-assessment with one specific CA in mind.

The fourth limitation pertained to the manipulation check. Whilst the levels of unfairness were successfully manipulated it was unexpected that the completely fair condition itself was significantly different than 0 as it should not elicit any flagging at all. This could be due to response bias (Wetzel et al., 2016). Even though the stimuli were randomly presented, this still means that the stimuli were completely fair or unfair six times in a row. People might not expect each answer from the CA to be completely fair or unfair. Thus, when the participants were doubtful, they might have coined it as fair if the previous stimuli were unfair and vice-versa. A future study could include conditions that are almost completely fair and almost completely unfair, to see if the participants flag more accurately.

A fifth limitation is the generalisability of the study's results. The study's sample mainly consisted of young adults. It is known that the elderly have barriers in using

information technologies due to unfamiliarity with the systems resulting in different usage than younger people (Fischer et al., 2014). It may be the case that they do not expect unfairness in CAs and, thus, would show different results for the second research question. For example, they might have a higher expected (pre) quality interaction, thus showing no difference between pre and post. A future study would have to include different age groups and populations.

A sixth limitation may have been that participants were primed to have a lower expected (pre) quality of interaction, which resulted in a high difference in the completely fair condition between the expected (pre) and experienced (post) quality of interaction. This can happen in studies where expectations are measured (Lau, n.d.). In this study, it could have happened, through the explanation of the study given at the start of the experiment (See Appendix C). Participants were told that they had to assess CA answers based on how reasonable the answers were. This may have primed them to have lower expectations of the CA. To ensure that this does not happen, future studies should only reveal the true aim after the participants have given their expectations.

Lastly, Bastiansen et al. (2022) considered the human likeness of the CAs an important factor in teasing out potential gender-based appearance differences. The current study used Fink's (2012) consideration of human likeness. Namely, a gendered profile picture and the use of pronouns. The limiting factor of the current study is that it is not known whether people considered the CAs human-like because of these additions. Furthermore, it is not known if participants were able to distinguish the sex of the CA at all, which also influences whether they perceived the CA as congruent. In future studies an additional question asking whether participants remembered the sex of the CA should be included. Finally, to even more firmly assess that appearance-based differences do not exist, in a future study, a set of CAs should be designed and assessed on human likeness by participants. Focussing on the language used by the CA, as suggested by de Sá Siqueira et al., (2023), could also be a fruitful venture point.

After the recommendations for future research based on the findings of the study, some additional recommendations are given that broaden the scope of the newly established foundation of the research area of unfair information in human-CA interactions. One direction could be looking at the influence of sexism-based unfair information in combination with the sex of the participant on the quality of interaction. It is known that women tend to be better than men at recognising sexist (misogynistic) remarks targeted at women (Drury & Kaiser, 2014). It could be that for women, in this situation, the quality of interaction is lower. To

research this, some of the other generated stimuli of the first study can be further validated and used to get a set of six solely based on unfair information containing sexism. Finally, the novel research area can be expanded by looking at the individual differences of participants. According to Piksa et al. (2022), both personality and susceptibility to certain cognitive processes play a large role in the uptake of unfair information, it might be that these factors also influence the perceived quality of interaction with unfair CAs.

References

- Arsham, H., & Lovric, M. (2011). Bartlett's Test. *International encyclopedia of statistical science*, 2, 20-23. https://www.researchgate.net/profile/Dr-Hossein-Arsham/publication/344731676_BARTLETT'S_TEST/links/5f8ce5cb299b1b53e324a18/BARTLETTS-TEST.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>
- Balakrishnan, V. (2022). COVID-19 and fake news dissemination among Malaysians – Motives and its sociodemographic correlates. *International Journal of Disaster Risk Reduction*, 73, 102900. <https://doi.org/10.1016/j.ijdrr.2022.102900>
- Bashkirova, A., & Krpan, D. (2024). Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. *Computers in Human Behavior: Artificial Humans*, 2(1), 100066. <https://doi.org/10.1016/j.chbah.2024.100066>
- 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>
- Berman, R., & Katona, Z. (2020). Curation Algorithms and Filter Bubbles in Social Networks. *Marketing Science*, 39(2), 296–316. <https://doi.org/10.1287/mksc.2019.1208>
- Beysolow II, T. (2018). What Is Natural Language Processing? In *Applied Natural Language Processing with Python* (pp. 1–12). Apress. https://doi.org/10.1007/978-1-4842-3733-5_1
- Blatná, D. (2006). Outliers in regression. *Trutnov*, 30, 1-6.

<https://statistika.vse.cz/konference/amse/PDF/Blatna.pdf>

- Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology & Marketing*, 38(7), 1052–1068. <https://doi.org/10.1002/mar.21480>
- Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Review of Medical Devices*, 18(sup1), 37–49. <https://doi.org/10.1080/17434440.2021.2013200>
- 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>
- Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide Web*, 675–684. <https://doi.org/10.1145/1963405.1963500>
- Chauhan, N. K., & Singh, K. (2018). A Review on Conventional Machine Learning vs Deep Learning. *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, 347–352. <https://doi.org/10.1109/GUCON.2018.8675097>
- Cooke, N. A. (2017). Posttruth, Truthiness, and Alternative Facts: Information Behavior and Critical Information Consumption for a New Age. *The Library Quarterly*, 87(3), 211–221. <https://doi.org/10.1086/692298>
- Dingler, T., Choudhury, A., & Kostakos, V. (2018). Biased Bots. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 1664–1668. <https://doi.org/10.1145/3267305.3274189>
- de Keersmaecker, J., & Roets, A. (2017). ‘Fake news’: Incorrect, but hard to correct. The role

- of cognitive ability on the impact of false information on social impressions. *Intelligence*, 65, 107–110. <https://doi.org/10.1016/j.intell.2017.10.005>
- de Sá Siqueira, M. A., Müller, B. C. N., & Bosse, T. (2023). When Do We Accept Mistakes from Chatbots? The Impact of Human-Like Communication on User Experience in Chatbots That Make Mistakes. *International Journal of Human–Computer Interaction*, 1–11. <https://doi.org/10.1080/10447318.2023.2175158>
- de Vreese, C. H., & Neijens, P. (2016). Measuring Media Exposure in a Changing Communications Environment. *Communication Methods and Measures*, 10(2–3), 69–80. <https://doi.org/10.1080/19312458.2016.1150441>
- Drury, B. J., & Kaiser, C. R. (2014). Allies against Sexism: The Role of Men in Confronting Sexism. *Journal of Social Issues*, 70(4), 637–652. <https://doi.org/10.1111/josi.12083>
- Fink, J. (2012). Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction (pp. 199–208). https://doi.org/10.1007/978-3-642-34103-8_20
- Fischer, S. H., David, D., Crotty, B. H., Dierks, M., & Safran, C. (2014). Acceptance and use of health information technology by community-dwelling elders. *International Journal of Medical Informatics*, 83(9), 624–635. <https://doi.org/10.1016/j.ijmedinf.2014.06.005>
- Følstad, A., & Brandtzaeg, P. B. (2020). Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience*, 5(1), 3. <https://doi.org/10.1007/s41233-020-00033-2>
- Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study (pp. 194–208). https://doi.org/10.1007/978-3-030-01437-7_16
- Hargittai, E., Fullerton, L., Menchen-Trevino, E., & Yates Thomas, K. (2010). Trust Online:

Young Adults' Evaluation of Web Content ° ESZTER HARGITTAI. In *International Journal of Communication* (Vol. 4). <http://ijoc.org>.

Gordon, C. (2023, February 2). ChatGPT is the fastest growing app in the history of web applications. *Forbes*. <https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/?sh=6c0c5f4a678c>

Hamby, A., Ecker, U., & Brinberg, D. (2020). How Stories in Memory Perpetuate the Continued Influence of False Information. *Journal of Consumer Psychology*, 30(2), 240–259. <https://doi.org/10.1002/jcpy.1135>

Hardmeier, C., Costa-jussà, M. R., Webster, K., Radford, W., & Blodgett, S. L. (2021). How to Write a Bias Statement: Recommendations for Submissions to the Workshop on Gender Bias in NLP. *ArXiv E-Prints*, arXiv:2104.03026. <https://doi.org/10.48550/arXiv.2104.03026>

International Organization for Standardization. (2018). *Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts* (ISO Standard No. 9241-11:2018). <https://www.iso.org/standard/63500.html>

International Organization for Standardization. (2019). *Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems* (ISO Standard No. 9241-210:2019). <https://www.iso.org/standard/77520.html>

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>

Jin, E., & Eastin, M. (2023). Gender Bias in Virtual Doctor Interactions: Gender Matching Effects of Chatbots and Users on Communication Satisfactions and Future Intentions to Use the Chatbot. *International Journal of Human–Computer Interaction*, 1–13. <https://doi.org/10.1080/10447318.2023.2279402>

Johnston, R., Jones, K., & Manley, D. (2018). Confounding and collinearity in regression

analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality & Quantity*, 52(4), 1957–1976.

<https://doi.org/10.1007/s11135-017-0584-6>

Johnson, D. R., & Creech, J. C. (1983). Ordinal Measures in Multiple Indicator Models: A Simulation Study of Categorization Error. *American Sociological Review*, 48(3), 398–407. <https://doi.org/10.2307/2095231>

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>

Kang, Y., Cappella, J., & Fishbein, M. (2006). The Attentional Mechanism of Message Sensation Value: Interaction between Message Sensation Value and Argument Quality on Message Effectiveness. *Communication Monographs*, 73(4), 351–378.

<https://doi.org/10.1080/03637750601024164>

Kaplan, A. D., Kessler, T. T., & Hancock, P. A. (2020). How Trust is Defined and its use in Human-Human and Human-Machine Interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1), 1150–1154.

<https://doi.org/10.1177/1071181320641275>

Kim, S., Lee, S., & Lee, J. (2021). Male, Female, or Robot?: Effects of Task Type and User Gender on Expected Gender of Chatbots. *Journal of Korea Multimedia Society*, 24(2), 320–327. <https://doi.org/10.9717/kmms.2020.24.2.320>

Lau, D. (n.d.) Priming Human-Centered Design Thinking for Enhanced Audience/User Experiences in Narrative Environments.

[https://www.researchgate.net/profile/Andreea-](https://www.researchgate.net/profile/Andreea-Felciuc/publication/376185674_Priming_Human-Centered_Design_Thinking_for_Enhanced_AudienceUser_Experiences_in_Narrative_Environments/links/656ce0b0b86a1d521b2e36a6/Priming-Human-Centered-Design-Thinking-for-Enhanced-Audience-User-Experiences-in-Narrative-Environments.pdf)

[Felciuc/publication/376185674_Priming_Human-](https://www.researchgate.net/profile/Andreea-Felciuc/publication/376185674_Priming_Human-Centered_Design_Thinking_for_Enhanced_AudienceUser_Experiences_in_Narrative_Environments/links/656ce0b0b86a1d521b2e36a6/Priming-Human-Centered-Design-Thinking-for-Enhanced-Audience-User-Experiences-in-Narrative-Environments.pdf)

[Centered_Design_Thinking_for_Enhanced_AudienceUser_Experiences_in_Narrative](https://www.researchgate.net/profile/Andreea-Felciuc/publication/376185674_Priming_Human-Centered_Design_Thinking_for_Enhanced_AudienceUser_Experiences_in_Narrative_Environments/links/656ce0b0b86a1d521b2e36a6/Priming-Human-Centered-Design-Thinking-for-Enhanced-Audience-User-Experiences-in-Narrative-Environments.pdf)

[_Environments/links/656ce0b0b86a1d521b2e36a6/Priming-Human-Centered-Design-Thinking-for-Enhanced-Audience-User-Experiences-in-Narrative-Environments.pdf](https://www.researchgate.net/profile/Andreea-Felciuc/publication/376185674_Priming_Human-Centered_Design_Thinking_for_Enhanced_AudienceUser_Experiences_in_Narrative_Environments/links/656ce0b0b86a1d521b2e36a6/Priming-Human-Centered-Design-Thinking-for-Enhanced-Audience-User-Experiences-in-Narrative-Environments.pdf)

- Li, Z. (2023). The Dark Side of ChatGPT: Legal and Ethical Challenges from Stochastic Parrots and Hallucination. *ArXiv E-Prints*, arXiv:2304.14347.
<https://doi.org/10.48550/arXiv.2304.14347>
- Lin, C.-C., Huang, A. Y. Q., & Yang, S. J. H. (2023). A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). *Sustainability*, 15(5), 4012. <https://doi.org/10.3390/su15054012>
- Mafra, M. G. S., Nunes, K., Rocha, S., Braz Junior, G., Silva, A., Viana, D., Silva, W., & Rivero, L. (2024). Proposing Usability-UX technologies for the design and evaluation of text-based chatbots. *Journal on Interactive Systems*, 15(1), 234–251.
<https://doi.org/10.5753/jis.2024.3856>
- Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59, 210–220.
<https://doi.org/10.1016/j.pragma.2013.07.012>
- 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>
- Modgil, S., Singh, R. K., Gupta, S., & Dennehy, D. (2021). A Confirmation Bias View on Social Media Induced Polarisation During Covid-19. *Information Systems Frontiers*, 26(2), 417–441. <https://doi.org/10.1007/s10796-021-10222-9>
- Muhammed T, S., & Mathew, S. K. (2022). The disaster of misinformation: a review of research in social media. In *International Journal of Data Science and Analytics* (Vol. 13, Issue 4, pp. 271–285). Springer Science and Business Media Deutschland GmbH.
<https://doi.org/10.1007/s41060-022-00311-6>
- Nicolescu, L., & Tudorache, M. T. (2022). Human-Computer Interaction in Customer Service: The Experience with AI Chatbots—A Systematic Literature Review. *Electronics*, 11(10), 1579. <https://doi.org/10.3390/electronics11101579>

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

<https://researchguides.njit.edu/evaluate/CRAAP>

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

<https://doi.org/10.1145/3495013>

Peterson, D. R., & Bry, B. H. (1980). Dimensions of perceived competence in professional psychology. *Professional Psychology*, 11(6), 965–971. <https://doi.org/10.1037/0735-7028.11.6.965>

Piksa, M., Noworyta, K., Piasecki, J., Gwiazdzinski, P., Gundersen, A. B., Kunst, J., & Rygula, R. (2022). Cognitive Processes and Personality Traits Underlying Four Phenotypes of Susceptibility to (Mis)Information. *Frontiers in Psychiatry*, 13.

<https://doi.org/10.3389/fpsy.2022.912397>

Poe. (n.d.). Poe: Fast, helpful AI chat. <https://poe.com/>

Qualtrics. (n.d.). Qualtrics XM: The leading experience management software. <https://www.qualtrics.com/>

Roetzel, P. G. (2019). Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research*, 12(2), 479–522. <https://doi.org/10.1007/s40685-018-0069-z>

Sanjeed, V. K., Kim, M. G., & Wang, C.-Y. (2020). Examining the Effect of Chatbot Gender and Gender Congruence between a Chatbot and a Customer in a Banking Context. *Journal of Korea Service Management Society*, 21(5), 46–73.

<https://doi.org/10.15706/jksms.2020.21.5.003>

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

<https://doi.org/10.1080/01621459.1968.10480932>

Shawar, B. A., & Atwell, E. (2007). Different measurements metrics to evaluate a chatbot system. *Proceedings of the Workshop on Bridging the Gap Academic and Industrial Research in Dialog Technologies - NAACL-HLT '07*, 89–96.

<https://doi.org/10.3115/1556328.1556341>

Siontis, K. C., Attia, Z. I., Asirvatham, S. J., & Friedman, P. A. (2024). ChatGPT hallucinating: can it get any more humanlike? *European Heart Journal*, 45(5), 321–

323. <https://doi.org/10.1093/eurheartj/ehad766>

Streicher, B., Frey, D., & Osswald, S. (2011). Forming Fairness Judgments: Why People Favor Unfair Information. In *Justice and Conflicts* (pp. 185–196). Springer Berlin Heidelberg.

https://doi.org/10.1007/978-3-642-19035-3_11

Stump, G. S., Husman, J., & Brem, S. K. (2012). The Nursing Student Self-Efficacy Scale.

Nursing Research, 61(3), 149–158. <https://doi.org/10.1097/NNR.0b013e318253a750>

Sullivan, G. M., & Artino, A. R., Jr (2013). Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4), 541–542.

<https://doi.org/10.4300/JGME-5-4-18>

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, 48(6), 1273–1296.

<https://doi.org/10.1007/s11165-016-9602-2>

Tomaskovic-Devey, D., & Mccann, C. (n.d.). *Who files discrimination charges? | Center for Employment Equity | UMASS*

Amherst. <https://www.umass.edu/employmentequity/who-files-discriminationcharges#:~:text=Race%20based%20charges%20are%20the,charges%20are%20much%20less%20common>.

University of Twente (n.d.). Master Biomedical Engineering.

<https://www.utwente.nl/en/education/master/programmes/biomedical-engineering/>

- Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How Reliable are Measurement Scales? External Factors with Indirect Influence on Reliability Estimators. *Procedia Economics and Finance*, 20, 679–686. [https://doi.org/10.1016/S2212-5671\(15\)00123-9](https://doi.org/10.1016/S2212-5671(15)00123-9)
- Valentino, N. A., Banks, A. J., Hutchings, V. L., & Davis, A. K. (2009). Selective Exposure in the Internet Age: The Interaction between Anxiety and Information Utility. *Political Psychology*, 30(4), 591–613. <https://doi.org/10.1111/j.1467-9221.2009.00716.x>
- van Berkel, E. Z. (2022). “Chatbots do not have a gender, they are just a piece of code.” *The effect of gender congruence between chatbot-product and chatbot-consumer on consumers’ perceived message trustworthiness and agent expertise*. [Master's thesis, Tilburg University]. <http://arno.uvt.nl/show.cgi?fid=157538>
- Van De Wouw, I. (2023). Evaluating hallucinations and repair in open-domain dialogue systems. [Bachelor's thesis, Radboud University]. Radboud Educational Repository. <https://theses.ubn.ru.nl/handle/123456789/15463>
- Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response Biases. In *The ITC International Handbook of Testing and Assessment* (pp. 349–363). Oxford University Press. <https://doi.org/10.1093/med:psych/9780199356942.003.0024>
- 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>
- 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 E-Prints*, arXiv:2309.01219. <https://doi.org/10.48550/arXiv.2309.01219>

Zhang, Y., Sun, J., Feng, L., Yao, C., Fan, M., Zhang, L., Wang, Q., Geng, X., & Rui, Y. (2024). See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–24. <https://doi.org/10.1145/3613904.3642545>

Zhu, Y., Zhang, J., Wu, J., & Liu, Y. (2022). AI is better when I'm sure: The influence of certainty of needs on consumers' acceptance of AI chatbots. *Journal of Business Research*, 150, 642–652. <https://doi.org/10.1016/j.jbusres.2022.06.044>

Appendix A: Survey Promotion Study 1

[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 biased; 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://utwentebbs.eu.qualtrics.com/jfe/form/SV_838gyHPMfMrCNdI

Appendix B: Generated Stimuli Study 1

This appendix contains the 32 stimuli as presented to the participants. The participants did not get to see whether the stimulus was fair or unfair, this is added for clarification. The stimuli were created around the hypothetical university of ACME but based on the Biomechanical Engineering Master at the UT, therefore any information directly taken from the site was quoted (UT, n.d.). This was also not visible to the participants to maintain the realistic aspect of questions about a hypothetical Master.

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.

Stimulus 3 Unfair

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

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

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

comprehensible. 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)” (UT, n.d., Master's structure section). 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)” (UT, n.d., Master's structure section). 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.

After each of these stimuli, a follow-up question regarding the confidence of the participant with regard to their given answer was asked. This question was the same for each

stimulus: How confident are you about your answer?

	Not confident at all	Slightly confident	Moderately confident	Very confident	Extremely confident
How confident are you about your answer?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix C: Introduction, Scenarios, Instructions and Disclaimers of Study 1 and 2

During both studies the participants were already, to some extent, informed about the goal of the study during the introduction. This enabled the participants to decide whether they wanted to partake in the experiment or not.

Introduction Study 1 – Stimulus Selection Study

Dear participant,

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

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

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

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

Contact details for further information:

Coordinator

Dr Simone Borsci

Dr Maryam Amir Haeri

Lucas Assen

Anna Bader

Nikola Markiewicz

Seán Verloop

Scenario Study 2 – Experimental Assessment of Interactions with (un)fair CAs

Dear participant,

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

The present study will take about 20-30 minutes to complete.

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

Your task will be to **determine whether the answer is "in line with your expectations"**: You can consider aspects such as, for instance, is the answer incomplete, unclear, poorly presented, or inappropriate etc. If the chatbot fails to meet your expectations, you will have the opportunity to explain why.

At the end of the survey, you will be asked to assess your overall experience with the Chatbot you interacted with.

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

There are no known major risks associated with the participation in this study, nevertheless, answers from conversational agents based on the knowledge they are provided with can be unpredictable, and even biased and unfair toward certain groups of people. Of course, this is a

simulation of a chatbot prototype, and if you identify such a behaviour, please do not take it personally and report it. All your personal information is confidential and can never lead back to you as an individual.

Contact details for further information:

Coordinator

Dr Simone Borsci

Dr Maryam Amir Haeri

Lucas Assen

Anna Bader

Nikola Markiewicz

Seán Verloop

After the introduction and during both studies 1 and 2, the participants were presented with an imaginary scenario and a set of instructions telling them what their goal was.

Scenario Study 1 – Stimulus Selection Study

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

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

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

- 1) Reasons for flagging: If you flag an answer as unfair or biased, we would also like you to report the reasons in a brief text;
- 2) Confidence in your decision to flag or not flag: Independent of whether you flagged an

answer as unfair/biased or not, we would like you to rate how confident you are that the answer is indeed fair/unbiased or unfair/biased 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).

Scenario Study 2 – Experimental Assessment of Interactions with (un)fair CAs

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

To do so, you are asked to interact with the university's chatbot by asking 6 of the most 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 the usability, trustworthiness, competence, and helpfulness of the chatbot.

Disclaimer of Both Studies

Finally, at the end of both studies, the participants were confronted with a disclaimer, which was the same for both studies:

Thank you for your participation in the study!

IMPORTANT

As a disclaimer, the unfair answers you have been exposed to during this study are examples of potentially biased statements that violate the Currency, Relevance, Authority, Accuracy, and Purpose test (https://en.wikipedia.org/wiki/CRAAP_test) these statements were generated using AI engines and tools provided with an unfair or biased set of information as knowledge to answer your questions. In the present study, we want to check the effect of such problematic knowledge on people's interaction with conversational agents.

If you would like to have additional information regarding the present study and future studies on this topic, please contact Dr Simone Borsci

Regards

their role as an advisor well							
The chatbots are usually useful for advice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The chatbots usually make it easy to find advice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The chatbots are usually helpful in finding advice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The chatbots are usually competent in giving advice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The information provided by the chatbots is usually credible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The information provided by the chatbots is usually factual	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The chatbots usually appear knowledgeable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I usually experience to get my questions answered by the chatbots	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

The scales of Bastiansen et al. (2022) and Borsci et al. (2022), with slight wording changes for the pre-assessment, when participants had not used CAs before, as presented to the participants:

When it comes to the **quality of interaction** of such types of systems [CAs], aspects like usability, helpfulness, competence and trust are considered very important. Please, rate your agreement with the following statements concerning **your expectations** in terms of the quality of interaction when you interact with these systems

Appendix E: Poe Chatbot Information

This appendix contains the information that was fed to the CAs, on which they based their answers. First, there were the prompts and general information that the CA uses to know how to behave and answer questions.:

Greeting

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

Context

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

###Main priority

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

Rules for answer the users

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

Keep your answer short, NO MORE THAN 150 WORDS

The [ADD] was replaced for each CA with its respective name and pronouns, dependent on the condition of the CA. Afterwards, general information regarding the subject i.e., the Biomedical engineering master, was fed to the CAs.

###General information about the MASTER BIOMEDICAL ENGINEERING AT THE UNIVERSITY OF ACME which was done by taking the information of the Master of

Biomedical Engineering at the University of Twente and redacting the UT's name to make sure that the CAs did not know it was about the UT (UT, n.d.).

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? (UT, n.d., Master Biomedical Engineering section).

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. (UT, n.d., Master Biomedical Engineering section)

CHOOSE ONE OF THE 4 (FOUR) SPECIALISATIONS:

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

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

1. Bioengineering,
2. MedRobotics,
3. InVitro & Medical Device
4. Signals & Systems (UT, n.d., Choose A Specialisation section).

Lastly, the improved stimuli based on the feedback of Study 1 were added to the CAs; it depended on the condition of the CA, whether the fair or unfair version was used. The stimuli were updated around the hypothetical university of ACME but based on the Biomechanical Engineering Master at the UT, therefore any information directly taken from the site was quoted (UT, n.d.). This was not visible to the participants to maintain the realistic aspect of questions about a hypothetical Master.

Stimulus 13 Fair

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.

Stimulus 13 Unfair

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.

Stimulus 16 Fair

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)” (UT, n.d., Master's structure section). 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

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)” (UT, n.d., Master’s structure section). 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

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.

Stimulus 17 Unfair

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.

Stimulus 20 Fair

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.

Stimulus 20 Unfair

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.

Stimulus 22 Fair

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.

Stimulus 22 Unfair

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.

Stimulus 25 Fair

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.

Stimulus 25 Unfair

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.

Table E1

A Table Providing the Name, Fairness, Appearance, and Access to Each of the CAs

CA	Fairness	Appearance	Link
OXXY she/her	Fair	female	https://poe.com/OXXY_she_her
OXXYA she/her	50% unfair	female	https://poe.com/OXXYA_she_her
OXXA she/her	Unfair	female	https://poe.com/OXXA_she_her
OXXI he/him	Fair	male	https://poe.com/OXXI_he_him
OXXIS he/him	50% unfair	male	https://poe.com/OXXIS_he_him
OXXIX he/him	Unfair	male	https://poe.com/OXXIX_he_him

Figure E1

Profile Icon of the Male CAs



Figure E2

Profile Icon of the Female CAs



Appendix F: R Code

```
---  
title: "Chatbot flagging behaviour"  
author: "Lucas et al."  
date: "`r Sys.Date()`"  
output: word_document  
---  
```{r setup, include=FALSE, echo=FALSE}  
require("knitr")
opts_knit$set(root.dir = "~/Desktop")
```  
  
```{r Packages and Library}  
#installing packages
install.packages("tidyverse")
install.packages("readxl")
install.packages("car")
install.packages("psych")
install.packages("dplyr")
install.packages("dgof")
install.packages("janitor")
install.packages("ggplot2")
install.packages("ggpubr")
install.packages("regclass")
install.packages("performance")
install.packages("tidyr")
install.packages("lme4")
install.packages("emmeans")
install.packages("effects")
install.packages("nlme")
library(tidyverse)
library(readxl)
library(car)
library(psych)
```

```
library(dplyr)
library(dgof)
library(janitor)
library(ggplot2)
library(ggpubr)
library(regclass)
library(performance)
library(tidyr)
library(lme4)
library(emmeans)
library(effects)
library(nlme)
...

```{r loading df}
#Importing the data
D_0 <- read_excel("FairChatBotDATACLEANED_V2.xlsx")
view(D_0)

#Descriptive statistics for the participant section
mean(D_0$Age)
summary(D_0$Age)
sd(D_0$Age)
...

```{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 and selecting data
```

```
D_1 <-
```

```
D_0 |>
```

```
select(Part = ID, Sex, Gender,
 Condition = condition_recoded,
 freq_use = `pre-experience frequency`,
 BUS1:COMP5,
 totBUSPRE:totalexperiencePRE,
 Flagging1:totFlag,
 BUS1post:COMP5post,
 totBUSPOST:D_totalexperience,
 NPS_NPS_GROUP:NPS) |>
```

```
mutate(Part = row_number()) |>
```

```
left_join(Conditions) |>
```

```
mutate(Sex_cong = (Sex == bot_sex))
```

```
...
```

```
```{r outlier analysis}
```

```
D_1$fairness = factor(D_1$fairness)
```

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

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

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

```
# outlier analysis taking into account the non-normality
```

```
M_outlier <- glm(totFlag_count ~ fairness, data = D_1, family = poisson (link = "identity"))
```

```
cooks_d <- cooks.distance(M_outlier)
```

```
# Plot Cook's Distance
```

```
plot(cooks_d, type="h", main="Cook's Distance", ylab="Cook's Distance", xlab="Participant")
```

```
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))
print(cooksd)
```

```{r descriptive statistics}
#appearance of the AI-driven-CA is renamed from nubers to male and female
D_4 <- D_1 %>% mutate(bot_sex = ifelse(bot_sex==1,"Male", "Female"))
D_4 <- D_4 %>% mutate(Sex_cong = ifelse(Sex_cong==TRUE,"Congruent",
"Incongruent"))

#the levels of fairness are renamed
D_4 <- D_4 %>% mutate(fairness = case_when(
  fairness == 0 ~ "Completely unfair",
  fairness == 0.5 ~ "50% unfair",
  fairness == 1 ~ "Completely fair",
  TRUE ~ "other"
))

#the fairness and appearance variables are turned into factors
D_4$fairness = factor(D_4$fairness)
D_4$bot_sex = factor(D_4$bot_sex)
D_4$Sex_cong = factor(D_4$Sex_cong)
D_4$fairness <- relevel(D_4$fairness, ref = "Completely fair")

#an overview of the total experience post and the delta total experience for each condition is
made.
D_10 <- D_1 %>% select(Condition, totalexperiencePOST, D_totexperience, bot_sex,
fairness)
D_11 <- D_10 %>%
  filter(Condition == "1")

```

```
summary(D_11)
```

```
D_12 <- D_10 %>%
```

```
  filter(Condition == "2")
```

```
summary(D_12)
```

```
D_13 <- D_10 %>%
```

```
  filter(Condition == "3")
```

```
summary(D_13)
```

```
D_14 <- D_10 %>%
```

```
  filter(Condition == "4")
```

```
summary(D_14)
```

```
D_15 <- D_10 %>%
```

```
  filter(Condition == "5")
```

```
summary(D_15)
```

```
D_16 <- D_10 %>%
```

```
  filter(Condition == "6")
```

```
summary(D_16)
```

```
D_17 <- D_4 %>% select(Sex_cong, D_totexperience)
```

```
D_18 <- D_17 %>%
```

```
  filter(Sex_cong == "Congruent")
```

```
summary(D_18)
```

```
D_19 <- D_17 %>%
```

```
  filter(Sex_cong == "Incongruent")
```

```
summary(D_19)
```

```
##box plots for total experience post and delta total experience are made which are separated  
for levels of fairness and appearance
```

```
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("Delta Quality of Interaction")+
ggtitle("The Delta Quality of Interaction Across the Conditions")
```

```
ggplot(D_4, aes(x = Sex_cong, y = D_totexperience, fill = Sex_cong))+
geom_boxplot(show.legend = FALSE)+
xlab("Chatbot Congruence")+
ylab("Delta Quality of Interaction")+
ggtitle("The Delta Quality of Interaction Over the CA's (In)congruence")
```

#Getting a summary of the individual quality of interaction variables for both the pre and post assessment across the different levels of fairness and appearance

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

```

summary_data_stdevs <- D_1 %>%
  group_by(fairness, bot_sex) %>%
  summarize(
    pre_trust = sd(tottrustPRE, na.rm = TRUE),
    post_trust = sd(tottrustPOST, na.rm = TRUE),
    pre_competence = sd(totcompetencePRE, na.rm = TRUE),
    post_competence = sd(totcompetencePOST, na.rm = TRUE),
    pre_helpfulness = sd(tohelpfulnessPRE, na.rm = TRUE),
    post_helpfulness = sd(tohelpfulnessPOST, na.rm = TRUE),
    pre_usability = sd(totBUSPRE, na.rm = TRUE),
    post_usability = sd(totBUSPOST, na.rm = TRUE),
    total_flagging = sd(totFlag, na.rm = TRUE),
    NPS = sd(NPS, na.rm = TRUE)
  )

summary_congruence <- D_4 %>%
  group_by(Sex_cong) %>%
  summarize(
    mdelta_qual = mean(D_totexperience, na.rm = TRUE),
    sdelta_qual = sd(D_totexperience, na.rm = TRUE)
  )
...
```{r testing scales}
#Cronbach's alpha for the scales is calculated to ensure their reliability. This action is
performed
#for each scale and for both the pre and post-usage. Furthermore, the raw alphas are stored in
a data frame and used in a bar plot with a line at 0.7 that functions as a threshold for the
acceptable levels for the Cronbach alpha.

#Pre assessment alpha
Alpha <- D_1 %>%
 select(BUS1:BUS9) %>%

```

```
psych::alpha(title = "pre assesment BUS", check.keys = TRUE)
alpha_PREBUS <-Alpha$total
```

```
Alpha2 <- D_1 %>%
 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 helpfulness", 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 assesment 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 helpfulness", check.keys = TRUE)
alpha_POSThelp <-Alpha7$total
```

```

Alpha8 <-D_1 %>%
 select(OMP1post:OMP5post) %>%
 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_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("Threshold", "Original alpha trust",
 "Original alpha competence", "Original alpha BUS", "Original alpha
helpfulness"))+
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
...

```{r parametric assumptions}
##normality totBUSPRE and POST
ggqqplot(D_1$totBUSPRE)
shapiro.test(D_1$totBUSPRE)

```

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

```
##normality tottrustPRE and POST
ggqqplot(D_1$tottrustPRE)
shapiro.test(D_1$tottrustPRE)
```

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

```
##normality tothelpfulness PRE and POST
ggqqplot(D_1$tothelpfulnessPRE)
shapiro.test(D_1$tothelpfulnessPRE)
```

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

```
##normality totcompetence PRE and POST
ggqqplot(D_1$totcompetencePRE)
shapiro.test(D_1$totcompetencePRE)
```

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

```
##normality total quality of interaction PRE and POST
ggqqplot(D_1$totalexperiencePRE)
shapiro.test(D_1$totalexperiencePRE)
```

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

```
##normality flagging
ggqqplot(D_1$totFlag)
```

```
shapiro.test(D_1$totFlag)
```

```
...
```

```
`` {r exploratory analyses}
```

```
#releveling of the factors to ensure that fairness, incongruent and male CA are is the reference categories
```

```
#To change the reference category to completely unfair change 'ref = "1"' to 'ref = "0"'
```

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

```
D_1$bot_sex = factor(D_1$bot_sex)
```

```
D_1$Sex_cong<-ifelse(D_1$Sex_cong, 1, 0)
```

```
D_1$Sex_cong = factor(D_1$Sex_cong)
```

```
#Manipulation check
```

```
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")
```

```
#Delta total flagging model
```

```
M_flag2 <- glm(totFlag_count ~ bot_sex + Sex_cong, data = D_1, family = poisson(link = "identity"))
```

```
summary(M_flag2)
```

```
confint(M_flag2)
```

```
VIF(M_flag2)
```

```
check_homogeneity(M_flag2, method = "bartlett")
```

```
#Delta total quality of interaction model
```

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

```
summary(M_total)
```

```
confint(M_total)
```

```
VIF(M_total)
```

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

Pairwise t tests to look at the difference between pre and post of the individual quality of interaction variables for the varying levels of manipulated (un)fairness.

```
D_7 <- D_1[D_1$fairness %in% c("1"),]
```

```
D_8 <- D_1[D_1$fairness %in% c("0.5"),]
```

```
D_9 <- D_1[D_1$fairness %in% c("0"),]
```

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

```
t.test(D_7$tottrustPRE, D_7$tottrustPOST, paired = TRUE)
```

```
t.test(D_7$totBUSPRE, D_7$totBUSPOST, paired = TRUE)
```

```
wilcox.test(D_7$totcompetencePRE, D_7$totcompetencePOST, paired = TRUE)
```

```
wilcox.test(D_7$tothelpfulnessPRE, D_7$tothelpfulnessPOST, paired = TRUE)
```

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

```
t.test(D_8$tottrustPRE, D_8$tottrustPOST, paired = TRUE)
```

```
t.test(D_8$totBUSPRE, D_8$totBUSPOST, paired = TRUE)
```

```
wilcox.test(D_8$totcompetencePRE, D_8$totcompetencePOST, paired = TRUE)
```

```
wilcox.test(D_8$tothelpfulnessPRE, D_8$tothelpfulnessPOST, paired = TRUE)
```

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

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

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

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

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

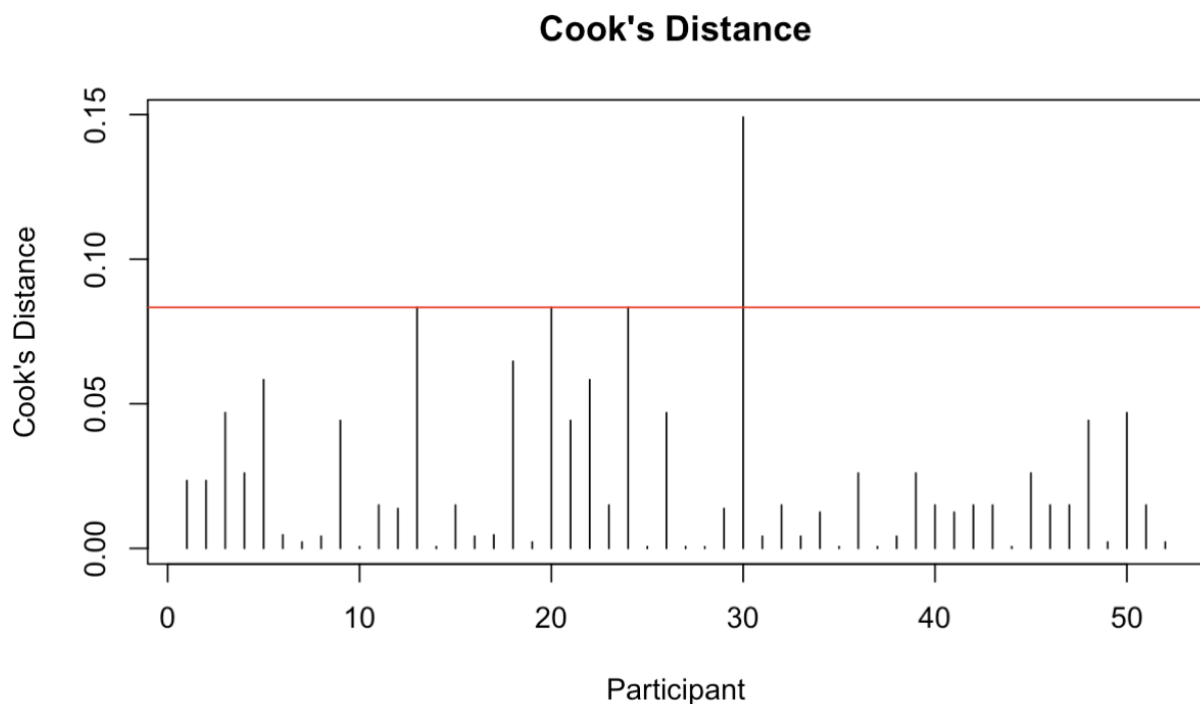
```
...
```

Appendix G: Outlier Analysis & Parametric Assumptions

For the outlier analysis, Cook's distance was used with a cut-off point of $4/n$. As the sample size was 52, it meant that any participant with a Cook's distance value over $D_i=.077$ was considered a significant outlier and removed. This resulted in the removal of 4 participants three of which were in the completely unfair condition and one of which was in the fair condition (See Figure G1). The participants (13, 20, 24, and 30) who surpassed the threshold had Cook's distance values of $D_i=.083$, $D_i=.083$, $D_i=.083$, and $D_i=.149$ respectively. Looking at the participants' behaviour, 13, 20, and 24 were in the completely unfair condition and flagged 0 stimuli as unfair it is thus reasonable to assume that they were indeed outliers. Participant 30 was in the fair condition and flagged three stimuli as unfair, thus again it was reasonable to assume that they were an outlier.

Figure G1

Plot Showing the Cook's Distance Values for Each Participant, With the Threshold Showing Which Participants are the Outliers



For the parametric assumption of normality, a Shapiro-Wilk test was performed on the main variables in the data. Based on the results it could be concluded that the variables, total competence pre and post, total helpfulness pre and post, and total flagging were non-normal as they had a p -value of $p<.05$ meaning that the alternative hypothesis i.e., the data is non-normally distributed, ought to be retained (See Table G1).

Table G1*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*
Quality of Interaction Pre	.97	.390
Quality of Interaction Post	.97	.310
Total Flagging	.90	< .001***

For the parametric assumption of the homogeneity of variances (homoscedasticity) across groups, the GLMs were subjected to a Bartlett test. The manipulation check showed a p-value of $p=.05$ (See Table G2). This means that this model is close to significance but does not show significant differences in the homogeneity of variances.

Table G2*Bartlett's Test on the Generalised Linear Models to Check for Homoscedasticity*

Models	p
Manipulation check Model	.050
Total Flagging Model	.698

Delta Δ Quality of
Interaction Model .481

Apart from looking at the homogeneity of variances of the models, they were also tested for multicollinearity. To do so the VIF between each independent variable in each model was calculated. As the manipulation check model only contained one independent variable, multicollinearity was of no concern. In the other models, no multicollinearity was found for any of the variables as all the VIFs were below 2.5 which is the threshold for mild multicollinearity (See Table G3).

Table G3

VIF Scores of the Independent Variables in the Generalised Linear Models to Test for Multicollinearity

GLM models	VIF Scores		
	Appearance of CA	CA congruence	Levels of manipulated (Un)fairness
Total Flagging Model	1.03(1)	1.03(1)	-
Delta Δ Quality of Interaction Model	1.04(1)	1.04(1)	1.02(2)