The Impact of GenAI-Disclosure on the Evaluation of Chatbot Conversations: Its Effects on User Perception in Successful Conversations and Conversational Failures.

Tijn A.J. de Kock

SNR: 2117704, ANR: 513320

Master's Thesis

Communication and Information Sciences

Specialization Communication and Cognition

Department Communication and Cognition

School of Humanities and Digital Sciences

Tilburg University, Tilburg

Supervisor: Dr. Christine Liebrecht

Second reader: Dr. Tom Lentz

January 2025

Technology statement

1. To paraphrase text from other sources, Quilbot was occasionally used.
2. To check spelling or grammar, Grammarly and ChatGPT was occasionally used.
3. To typeset the given text, no tools were used.
4. To generate part of the text, no tools were used.
5. Perplexity AI and Scispace were used only as an aid in finding relevant sources of information.

By submitting my thesis for assessment, I hereby confirm that:

-The thesis is my own intellectual property and that ideas as well as language from other sources have been properly cited. All quotes and source information have been properly identifiable as such

-I have disclosed any technology that I have used in the writing process.

Table of content

Abstract

Generative AI (GenAI) chatbots have been gaining traction in the customer service area because of their advanced capabilities. However, despite their advancements, the possibility of errors like hallucinations raises questions about their reliability and users' trust. This study examined the effects of disclosing the use of GenAI in a chatbot's introduction message on the user's trust and perceived expertise. Moreover, it examines the effects of a chatbot's conversational failure on trust and the violations of expectations. Participants engaged in an experiment in which trust was measured at three points: before the introduction message, after reading the introduction message, and after interacting with the chatbot. The findings indicate that GenAI-disclosure does not affect trust, perceived expertise, and expectancy violations. No significant difference was found on these variables between disclosure conditions. Trust did increase significantly after the introduction message and increased further after successful interactions, however, it dropped significantly after errors occurred. The results show that while GenAI-disclosure does not increase nor harm user trust, organizations should focus on mitigating the negative effects of conversational failures to maintain a good user experience. The results provide valuable insights into how trust develops during chatbot interactions and provide recommendations for improving transparency and the user experience.

Keywords: chatbot, GenAI, disclosure, trust, expertise, expectancy violations, conversational failure, user experience

The Impact of GenAI-Disclosure on the Evaluation of Chatbot Conversations: Its Effects on User Perception in Successful Conversations and Conversational Failures

Chatbots, defined as automated computer programs communicating with humans in natural language (Chakraborty et al., 2023), have been increasingly used for customer support in recent years (Følstad et al., 2018). They use Natural Language Processing (NLP) to recognize human language, making them able to 'understand' the user's questions and respond in an automated way, often based on a predefined set of rules (Adamopoulou & Moussiades, 2020a). These rule-based chatbots have provided businesses with many possible advantages such as 24/7 customer support, reduced response times, and savings in labor and money (Chhabra et al., 2024; Hsu & Lin, 2023). However, they have significant limitations regarding flexibility and handling user input errors (e.g. questions with spelling mistakes or different wording than what the chatbot is programmed to recognize).

Meanwhile, a new type of chatbot using Generative AI (GenAI), can generate responses by collecting relevant data using deep learning (Chakraborty et al., 2023). This makes it able to answer in a better-suited and more flexible way than rule-based chatbots (Adamopoulou & Moussiades, 2020a). Not surprisingly, increasingly more companies are now interested in having their customer service chatbots powered by GenAI (Maedche et al., 2019). As many businesses see the advantages and are shifting towards GenAI in their customer service chatbots, the question emerges: what are their users' perceptions of this technology, and are they comfortable relying on answers generated by AI? Moreover, the European Commission has been discussing legal regulations to disclose the identity of chatbots to create higher transparency, making firms obligated to reveal the chatbot's non-human identity. Disclosing a chatbot's identity will likely be mandatory and this influences user's perceptions of chatbots (Mozafari et al., 2021). Considering this, it is critical to understand to what extent users' expectations are different with a GenAI chatbot compared to a traditional rule-based one and if it is advantageous for companies to use AI-disclose alongside chatbot-disclosure.

Managing users' expectations is crucial in chatbot adoption, as many people are hesitant about interacting with them. Chatbots regularly trigger dissatisfaction and even frustration among users. Findings of a recent survey showed that 53% of users thought that chatbots were "not effective" or

only “somewhat effective” (Hsu & Lin, 2023). Research shows that emphasizing a chatbot’s expertise can heighten users’ trust and expectations for their interaction with a chatbot. When a chatbot is presented as an expert, the user’s expectations, as well as the satisfaction after a successful conversation will increase (Liebrecht et al., 2024). Expertise can be emphasized by stating one’s experience or training in a specific field of knowledge (Farrington-Darby & Wilson, 2005), as in the study of Liebrecht et al. (e.g. ‘I am an expert in … I have successfully helped hundreds of customers.’). Expertise can also be emphasized by identifying one’s source of information (e.g. ‘I am an expert powered by ChatGPT) (Birnbaum & Stegner, 1979), which is different between rule-based chatbots and GenAI chatbots. Given this, disclosing GenAI can be seen as a way of disclosing the chatbot’s source of information and possibly results in altered perceptions of expertise and trust.

A key question is how the disclosure of GenAI influences users’ perception of the chatbot’s expertise and trust. In recent years people across industries have increasingly encountered GenAI with tools such as ChatGPT (Khan et al., 2024). Evaluations of these programs are generally positive as previous studies showed that 85% of ChatGPT-powered conversational interfaces were rated with high levels of satisfaction by their users (Sakirin & Said, 2023). This suggests that expectations for chatbots could be further elevated if GenAI is mentioned alongside expertise.

However, GenAI chatbots are not always free of errors. Among others, problems in interpreting user input can happen, just like with rule-based models. Misunderstandings, where a wrong interpretation is given to the user’s expressions (e,g, User: "My order arrived damaged." Chatbot: "Your order status is: Delivered. Anything else I can help with?"), are one of the common errors in responses from chatbots (Bohus & Rudnicky, 2008; Liebrecht et al., 2024). Sometimes, GenAI chatbots even generate false information, known as hallucinations (Alkaissi & McFarlane, 2023). These errors remain hard to eliminate and might create distrust.

When users encounter errors in GenAI chatbots, their high expectations may lead to expectancy violations: a negative reaction that occurs due to high pre-encounter expectations and poor post-usage performance (Crolic et al., 2021). This hurts the user’s satisfaction of the interaction and reduces the overall evaluation of the chatbot (Brendel et al., 2023; Cai et al., 2024). Users’ previous experiences with these errors of AI-generated content might diminish their trust in the AI-generated

responses and lower perceptions of expertise. Previous studies have demonstrated various negative effects of expectancy violations after chatbot errors (Brendel et al., 2023), but the question is whether this is different when GenAI is disclosed in advance and expectations about the chatbot are altered.

How these expectations will be altered remains unclear given the generally positive evaluation of GenAI among users, but the possibility for errors. More insight into the expectations and perceptions towards GenAI in chatbots is needed to know what the effects of disclosing GenAI are on the interplay between the user's perceived expertise, trust, and expectancy violations in successful conversations and conversations with errors. While previous studies have already explored some of the general effects of disclosing AI, little emphasis is placed on GenAI and its comparison to rule-based chatbots. Additionally, the effects on trust, perceived expertise, and expectancy violations after conversational failure remain unstudied. As businesses are increasingly showing interest in implementing GenAI in their customer support chatbots and regulating laws might obligate the disclosure of it, knowing these differences in expectations and evaluations is relevant in helping companies understand and improve customers' satisfaction and adoption of these GenAI chatbots.

This study links these concepts of expertise, trust, and expectancy violations to the disclosure of GenAI by comparing two distinct framing approaches to the chatbot's introduction -'chatbot based on ChatGPT' versus 'chatbot'- and explores how these differences influence user perception in an advantageous or disadvantageous way. With this, the research question is as follows: 'To what extent does disclosing the use of GenAI in a customer service chatbot influence users' perceptions of expertise, trust, and expectancy violations?'

Theoretical framework

Chatbots and Artificial Intelligence

Chatbots are able to mimic human conversation, which can serve purely entertaining purposes but can also be useful in many more applications such as information retrieval, education, healthcare, and business (Shawar & Atwell, 2007). The rapid growth of interest started particularly after 2016, with many chatbots being developed for industrial solutions such as customer service (Adamopoulou & Moussiades, 2020a). The ability to handle multiple users simultaneously and the reduction in

customer service costs are some of the main reasons why chatbots have been growing in popularity in businesses (Hsu & Lin, 2023).

Artificial Intelligence can be defined as "systems that display intelligent behavior by analyzing their environment and taking actions, with some degree of autonomy, to achieve specific goals." (Sheikh et al., 2023). AI has led to the development of a wide range of machine learning models, which can serve many applications. One of which is Natural Language Processing (NLP), which enables machines to understand, interpret, and respond to human language (Dam et al., 2024; Khanna et al., 2015). However, the degree to which chatbots use this technology varies across models. Although there are different ways of categorizing chatbots, a distinction can be made between rule-based and generative models.

Most of the first chatbots were built with a rule-based architecture (Esfandiari et al., 2023). These rule-based model chatbots respond by using a fixed predefined set of rules which are mostly based on the recognition of the user's input using simple Natural language Understanding (NLU) - a subset of NLP - techniques like pattern matching. They do this without the creation of any new text answers. The source of information used in the response of a rule-based chatbot is pre-written answers created by humans that are organized into a rule database so that the system can link questions to certain answers. A more extensive rule database allows the chatbot to respond to a wider variety of user input. However, the answer provided is always a pre-written text that matches best with the question. The upside to this is that there is full control over the response that the chatbot provides, limiting false information. The downsides of these rule-based chatbots are their lack of variation in generated answers, the fact that they are limitedly resilient to a user's grammatical and spelling mistakes, and their inability to take earlier parts of the conversation into consideration in their response (Adamopoulou & Moussiades, 2020a). Because of their pre-written answers without the possibility of adaptation to the specific user's input, rule-based chatbots are relatively limited in the range of questions they can handle and their ways of responding. These limitations in rule-based chatbots can lead to miscommunications and frustrations among their users (Zhang et al., 2024).

The generative model (GenAI chatbot) creates answers in a more advanced way than the rule-based model. These chatbots are pre-trained with deep learning techniques along with machine

learning algorithms to possess Large Language Models (LLM) as a large source for their information (Adamopoulou & Moussiades, 2020a; Wang et al., 2023). It uses more advanced NLU techniques to understand user input in a more complex way. It also uses Natural Language Generation (NLG), yet another subset of NLP to generate its own answers. After receiving input from the user, the generation of a response can be divided into two parts: content planning and content realization. Content planning entails selecting, prioritizing, and organizing key information based on the perceived user needs. Content realization concerns structuring this content into sentences, generating references, and using discourse cues (Adamopoulou & Moussiades, 2020b). This is how a GenAI chatbot generates its answer using NLG.

The differences between rule-based chatbots and GenAI chatbots are as follows. Rule-based chatbots use simpler techniques mainly for basic language processing and pattern matching, while GenAI chatbots use more advanced NLP techniques to understand and process natural language, allowing them to handle more complex and varied language structures and improve their language (Esfandiari et al., 2023). GenAI chatbots use sophisticated NLU to extract meaning, context, and intent from user input making it better able to understand nuances, ambiguities, and contexts and use the history of the conversation. Meanwhile, rule-based chatbots do not really have understanding capabilities, mainly focusing on keyword matching and predefined patterns making them struggle with understanding context, variations, and nuances and unable to account for chat history. Finally, GenAI chatbots use NLG to generate human-like responses, allowing for variety and personalized, context-appropriate content, but also increasing the possibility for falsely constructed information. Contrarily, rule-based chatbots use pre-written responses, making the answers fully controllable but very limited in variation and adaptation to context (Adamopoulou & Moussiades, 2020b; Khurana et al., 2022). In general, GenAI chatbots provide various benefits in terms of functionality over rule-based chatbots, which is why increasingly more companies are interested in implementing a GenAI chatbot for their customer service.

Conversational Failure

Improvements in GenAI chatbots over rule-based chatbots are believed to help in the reduction of conversational failures - situations in which the user does not receive the information that it

aspires to. Although this is generally the case, conversational failures cannot be completely eliminated and can result in undesired responses from the user's viewpoint. Conversational failures can be categorized into non-understanding, misunderstandings, and hallucinations (Alkaissi & McFarlane, 2023; Liebrecht et al., 2024). Non-understandings are conversational failures where the system cannot give a valuable interpretation of the input from the user. The system recognizes this and might not respond or respond likewise (e.g. 'Sorry, I don't understand.'). Non-understandings are therefore relatively easy to recognize for the user. Misunderstandings, however, occur when a wrong interpretation is given to the user input (Liebrecht et al., 2024). The chatbot responds with a truthful answer, but it is not an answer to the question that the user intends to get an answer to. Misunderstandings are slightly harder to recognize for chatbot users than non-understandings. Lastly, hallucinations occur when a chatbot generates an answer that seems truthful, and to answer the right question, but in fact contains false information (Alkaissi & McFarlane, 2023). These failures are very hard to recognize for users as one has to know the correct information to spot false information.

There are several possible causes for conversational failures such as questions outside of the chatbot's application domain (Bohus & Rudnicky, 2008), questions with too much complexity or ambiguity, or unrecognized formulation (Dzikovska et al., 2009). How a chatbot handles these situations is dependent on the type of chatbot. Non-understandings are most prevalent in rule-based chatbots because of their limited NLU capabilities. The input cannot be linked to a certain pattern and the system is not able to give a valuable interpretation. Misunderstandings happen in rule-based chatbots because questions can be linked to the wrong answers. However, they are also prevalent in GenAI chatbots because of their ability to interpret and extract meaning out of input, which can go wrong due to the causes mentioned above. Hallucinations happen only in GenAI chatbots, as responses of rule-based chatbots are pre-written and the info is thus controlled, while GenAI chatbots' answer generation is less controllable.

The focus in this study will be on misunderstandings. The reasoning for this is that, to compare rule-based chatbots to GenAI chatbots, the conversational failure must be able to be linked to both models while it also has to be detectable. Non-understandings are detectable but mostly prevalent in rule-based chatbots while hallucinations are inherently impossible to detect and only prevalent in

GenAI chatbots. Examining the impact of misunderstanding is therefore believed to be most relevant and effective in the context of this study. As misunderstandings are detectable for the user and prevalent in both rule-based and GenAI chatbots, implementing them in the conversation allows for a realistic comparison between the two models in examining the effects of conversational failures on user perceptions.

Examining the impact of conversational failures is critical considering the frequency of occurrence and their effects. According to a survey from the banking industry, nearly 75% of customers report that chatbots often fail to offer accurate answers resulting in dissatisfaction with the interaction in 80% of customers. (Haupt et al., 2023). Chatbot mistakes negatively affect many aspects including the user's trust, and perceived competence of the agent (Gu et al., 2024; Toader et al., 2019). These errors are frequently complained about by users negatively influence how people in general perceive and trust chatbots, slowing down the adoption of this technology. As conversational failures cannot be completely prevented due to restrictions in technology and the unpredictability of user input (De Sá Siqueira et al., 2023), it is important to focus on mitigating the negative effects. One way of mitigating these negative effects of conversational failures is by managing users' expectations.

Social Cues in Introduction Messages

The chatbot's introduction message is an effective way to steer its users' expectations, as it is essentially the first step of the customer's communication journey with it (Van Hooijdonk et al., 2023). A well-constructed introduction message can increase engagement (Kull et al., 2021), and focusing on the right cues in the introduction message is crucial in reaching desired outcomes. These cues are so-called social cues -"biologically and physically determined features salient to observers because of their potential as channels of useful information" (Fiore et al., 2013, p.2)- and can be divided, among others, into identity cues and competence cues (Van Hooijdonk et al., 2023).

A disclosure is a way of transmitting social cues. One can disclose a chatbot's identity (chatbot disclosure), whether or not you disclose the non-humanness of the chatbot in the first place, and one can further disclose its source, as with GenAI-disclosure (e.g. 'This chatbot is powered by ChatGPT.'). Chatbot disclosure transmits identity cues and is an important topic in the context of chatbots' introduction messages as the idea of legal regulations that obligate the disclosure of the non-human

identity of chatbots is gaining traction worldwide. California has already formed a ‘bot bill’ (fining undisclosed chatbots) and the European Commission has also been having conversations about this topic with the idea of creating higher transparency to reduce deception and exploitation by service providers as it is increasingly difficult for users to know if they are interacting with a human or a computer (Mozafari et al., 2021). The disclosure of a chatbot’s identity can have significant effects on the user’s expectations and post-evaluations of the interaction with and content from a chatbot (Lim & Schmälzle, 2024). Previous research has shown negative effects of disclosing the non-human identity of a chatbot, mediated by decreased humanness (Lim & Schmälze, 2024). As the disclosure of a chatbot’s identity may be inevitable because of legal restrictions, how it is disclosed in order to reach positive outcomes increases in importance.

To strategically present social cues that present the chatbot in a favorable way, one can signal competence cues. Signalling only these preferable cues like competence cues (a process called selective self-presentation), can build trust (Li et al., 2024; Mazafari et al., 2021). Signalling expertise can enhance competence cues, which can be done by disclosing ones source of information (Birnbaum & Stegner, 1979). As the source of information is different for GenAI chatbots (employing a LLM) compared to rule based chatbots (employing a rule database), disclosing GenAI can perhaps alter perceived expertise and the overall competence cues, which can influence a user’s trust levels. The primary goal of a chatbot’s introduction message is to focus on the right cues to effectively manage user expectations. In this study, this involves applying selective self-presentation mechanisms to frame the chatbot in a way that optimizes users’ perceptions of expertise and trust.

Expertise

Expertise can be described as a high level of knowledge, skills, or abilities in activities or tasks within a certain domain. High levels of expertise are linked to several characteristics such as extensive and up-to-date content knowledge, high quality of performance, and ability to communicate (Farrington-Darby & Wilson, 2005). These characteristics are important for productivity, which is the primary motivator for chatbot users (Adamopoulou & Moussiades, 2020b). It is thus important that chatbot users perceive high levels of expertise from the chatbot that they are about to interact with. In terms of social cues (Fiore et al., 2013), mentioning expertise can be used in a chatbot’s introduction

message displaying competence cues as a way of selective self-presentation.

High levels of expertise-perception have several advantages for the usage and evaluation of chatbots. A study done by Liebrecht et al. (2024) showed that when a chatbot is introduced as an expert (e.g. “I am an expert in booking flights and have successfully helped hundreds of customers”), the users’ expectations before interaction and evaluation after a successful interaction were higher and more positive compared to the same interaction but with the chatbot being introduced as ‘in training’. Another positive effect of perceived expertise is that when mentioning the expertise of a chatbot in its introduction message, the negative effects of disclosing its non-human identity, are mitigated and can even attain trust levels that are equivalent to those of undisclosed agents (Mozafari et al., 2021).

What remains unexplored yet is whether the disclosure of GenAI alongside the chatbot’s identity disclosure can be seen as a form of expertise. Although the effects on perceived expertise are unknown, previous studies have already investigated the influence of agent framing (as an intelligent entity versus as a machine) in a chatbot’s introduction message on other aspects like social presence or perceived social intelligence, showing no significant main effects (Araujo, 2018; Xu et al., 2023). However, the framing in these studies did not focus on GenAI and further investigation of different variations of framing was recommended. Xu et al. also investigated transparency (giving an explanation in simple language about how the chatbot works) which did significantly increase perceived social intelligence. Moreover, it is known that transparency in the decision-making process enhances perceptions of competence (Wang & Benbasat, 2007). As with the transparency manipulation in the study of Xu et al., GenAI-disclosure could also provide the user a sense of explanation of the workings of the given chatbot and increase transparency. With this, GenAI-disclosure could signal competence cues, although further research is needed to investigate this.

By increasing transparency, GenAI-disclosure in a chatbot’s introduction message could be a way of signalling competence, increasing the user’s perceived expertise. Moreover, combining the technological advancements of GenAI chatbots over rule-based chatbots and the increased usage and awareness of GenAI by the general population, it is reasonable to assume that more people are aware of these advancements and that GenAI-disclosure has a positive effect on perceived expertise compared to only chatbot disclosure. Based on this reasoning, H1 is formulated.

H1: GenAI-disclosure enhances the levels of perceived expertise compared to chatbot disclosure.

Trust

While productivity remains a primary motivator for chatbot users, trust in a chatbot is equally crucial for a user's willingness to engage with it (Følstad et al., 2018). Trust is 'an expectancy held by an individual that the statement of another individual can be relied upon' (Rotter, 1967, p. 651). It is essential for adopting automation technologies (e.g. chatbots), and a lack of trust can result in decreased reliance on these systems (Lee & See, 2004). Demonstrating competence and expertise is claimed to be the most influential element for people's trust in chatbots (Mozafari et al., 2021). As the disclosure of GenAI is hypnotized to enhance users' expertise perceptions, trust levels could be expected to enhance too.

However, studies that investigated expertise disclosure in chatbots on trust are limited and as trust in AI is dependent on several more factors besides expertise (Li et al., 2024), the direct effect of GenAI-disclosure needs to be examined. There are some studies that investigated the effects of AI-disclosure, which reported negative effects on the evaluation of messages and fairness of decisions (Lim & Schmälze, 2024; Newman et al., 2020). These studies indicate that the source of a message can bias people's assessments of conversations and content. However, these studies compared AI-disclosure to non-disclosure, which led to a decrease in perceived humanness - another influential factor of trust (Li et al., 2024). To make more direct claims about the perceptions towards GenAI and its impact on trust, the non-humanness of the chatbot must already be clear to the user.

Measuring effects of AI-disclosure alongside the chatbot's identity disclosure on user evaluations could examine this further. Previous studies have started to explore this. A study by Aujaro (2018) compared two ways of agency framing - an intelligent frame ("a virtual agent powered by artificial intelligence (AI) which uses machine learning and AI technology to engage in conversations automatically") versus a neutral frame ("*virtual agent*") – on users' perceptions of humanness and social presence. There was no main effect of framing on perceived humanness or social presence but when other social cues (human-like language and name) were left out, the intelligent frame further lowered some levels of perceived humanness. The intelligent frame significantly lowered social presence when social cues were left out, however, it significantly improved social presence when these

social cues were implemented. This suggests that a chatbot's social cues are more influential users' evaluations than framing, but using an intelligent frame could enhance the effects of these social cues. The current study will adopt the human-like language and name disclosure as social cues from Aujaro but enrich it by adding expertise disclosure as social cue expressing competence, which is known to increase trust levels (Li et al., 2024). Aujaro suggested further research into different ways of framing, which this current study will do by using 'ChatGPT' to link findings to GenAI specifically. Moreover, Aujaro's study was conducted in 2018, before the general public became aware of the concept of (Gen)AI (Khan et al., 2024), so it is possible that its disclosure can have different effects today.

In the specific context of GenAI chatbots versus rule-based chatbots, experience with Generative AI could also be an influential factor for trust. Previous research has shown that past experiences with chatbots generally enhance factors like social presence and use intention which can be associated with trust (Min et al., 2021). This effect could also be prevalent in past experiences with GenAI. However, the aspect of hallucinations are unique to GenAI (chatbots). Ahmad et al. (2023) studied ChatGPT usage among students and showed that many have had encounters with hallucinations. Previous research showed that experience with hallucinations undermines trust (Kim, 2024), and when a conversational agent fails to meet expectations a couple of times (generally between 2 and 6 times), expectations are set (Luger & Sellen, 2016). Contrarily, ChatGPT's popularity is growing indicating positive experiences among users (Aydın & Karaarslan, 2023). Experience with GenAI tools like ChatGPT could have either positive or negative effects on trust and might be dependent on personal experiences.

To conclude, trust in AI is a multifaceted issue and it is not entirely clear how AI-disclose will affect trust in chatbots. Some research has been done on the effects of AI-disclosure compared to no disclosure, and the effect of framing on particular aspects related to trust. However, as trust can be dependent on many factors and different ways of framing are unexplored, direct measurements of the effect of GenAI-disclosure on trust are crucial. As currently available information slightly leans toward the positive effects of GenAI-disclosure on trust, the following hypothesis is formed:

H2: GenAI-disclosure increases trust levels compared to chatbot disclosure.

Expectancy Violations

To fully explore users' trust, it is important to also measure it after the conversation with the chatbot. A chatbot's introduction message can shape expectations but the question is whether these expectations can be fulfilled. High expectations shaped by increased perceptions of expertise and trust are positive for users' intentions to interact with chatbots. When the subsequent conversation is successful and these expectations are confirmed, it will be positive for user satisfaction. However, when a conversational failure occurs, dissatisfaction arises (Qin, 2023) and these high expectations might not be met. This leads to negative disconfirmation known as expectancy violations. Monitoring and minimizing expectancy violations is a crucial factor for organizations implementing customer service chatbots as they are shown to negatively impact evaluations of chatbots and the related company and even evoke feelings of betrayal (Cai et al., 2024c; Crolic et al., 2021; Saeed et al., 2024). Given this, expectancy violations caused by a conversational failure could lower a user's trust.

Expectancy violations can occur because of high pre-set expectations or poor post-usage performance (Crolic et al., 2021). Among other factors, user expectations can be raised by high perceptions of expertise and an increase in trust. As hypothesized, the disclosure of GenAI could likely increase perceived expertise, raising expectations about chatbot performance. When a chatbot's behavior does not match its expertise label, it creates expectancy violations (Rheu et al., 2024). Similar to the study of Liebrecht et al. (2024) where the positive effects of expertise disclosure disappeared when a conversational failure occurred. The expected effects of AI-disclosure on trust levels remain more ambiguous but are also hypothesized to increase trust levels. This would likely raise expectations and therefore increase expectancy violations after conversational failure even further. However, if GenAI-disclosure lowers trust levels and makes users more aware of the possibility of conversational failures (i.e. hallucinations), expectancy violations after encountering such an error will likely decrease because these errors are accounted for and fit within the user's expectations. If this is the case, expectations will likely not be violated as much but whether the increase in expertise or the decrease in trust has more effect on expectancy violations is unknown. While perceived expertise may lead to higher expectations, trust may be more influential on the effects

of violations of these expectations.

To deeper understand the expectations people have regarding GenAI, expectancy violations will be measured both after successful conversations and after conversational failures. This approach creates insight into whether people actually expect conversational failures from generative AI. Compared to several years ago, people have gained increased experience with chatbots and GenAI, which could be positive or negative. Overall, it is hypothesized that GenAI-disclosure will increase expertise and trust and that this can increase dissatisfaction after encountering a conversational failure. This leads to the formation of the third hypothesis:

H3: GenAI-disclosure increases expectancy violations after conversational failure compared to chatbot disclosure.

In order to fully understand the user's journey in terms of trust in a chatbot and to understand the impact of the introduction message and the successfulness of the conversation, trust will in total be measured across three moments. Firstly, trust will be measured before seeing the introduction message of the chatbot, serving as a baseline measurement for the user's trust levels in chatbots without any manipulation. Secondly, as a chatbot's introduction message can shape expectations, trust will be measured alongside expertise after seeing the introduction message to gain knowledge on the effects of GenAI-disclosure. Thirdly, trust will be measured after interacting with the chatbot, comparing successful conversations with conversational failures. This gives an indication of how much effect conversational failures have on trust levels and if this can be altered by disclosing GenAI. This links trust in GenAI to the expectation of its hallucinations and conversational failures in general. With this, the final hypothesis is formed:

H4: GenAI-disclosure decreases trust levels after conversational failure compared to chatbot disclosure.

Method

Experimental Design

To investigate the effects of disclosing GenAI in a chatbot's introduction message on the

expectations and evaluations of the successful and unsuccessful interaction, an experiment using a 2x2 between-subjects design was conducted. The first independent variable was the disclosure in the introduction message, where the chatbot was disclosed either as a 'chatbot based on ChatGPT' or merely as a 'chatbot'. The second independent variable was the conversational failure, whether the conversation was completed successfully or resulted in a conversational outcome. Participants were randomly assigned to one of the four distinct chatbot versions created for this study. Dependent variables - trust, perceived expertise, and expectancy violations - were measured through a survey administered via Qualtrics.

Participants

The participants were recruited using a convenience sampling approach. The researcher distributed the experiment primarily through social networking platforms within their personal and professional network. Additionally, participants were recruited using a snowball sampling method and the online platform 'SurveyCircle'. These sampling methods allowed for reaching a total amount of 203 respondents who finished the experiment. The GenAI-failure, GenAI-disclosure-success, undisclosed-failure, and undisclosed-success conditions had 53, 53, 47, and 50 participants respectively. A total of 76 men and 126 women participated and one participant preferred not to say their gender. The average age of participants was 29 years old with a minimum of 17 and a maximum of 78 and education levels varied between 'less than high school' and 'master's degree'. Randomization on these aspects was secured. The χ^2 test of association revealed no statistically significant differences in age between the conditions ($\chi 2(3) = 1.06$, $p = .79$). The chi-square test showed no significant association between the participants' gender and the conditions, $\chi^2(6) = 9.54$, $p = .15$, nor the participants education level and the conditions, $\chi^2(12) = 13.70$, $p = .32$. The results from the one-way ANOVA revealed no significant difference in age between conditions, $F(3,199) = 1.00$, $P = .39$.

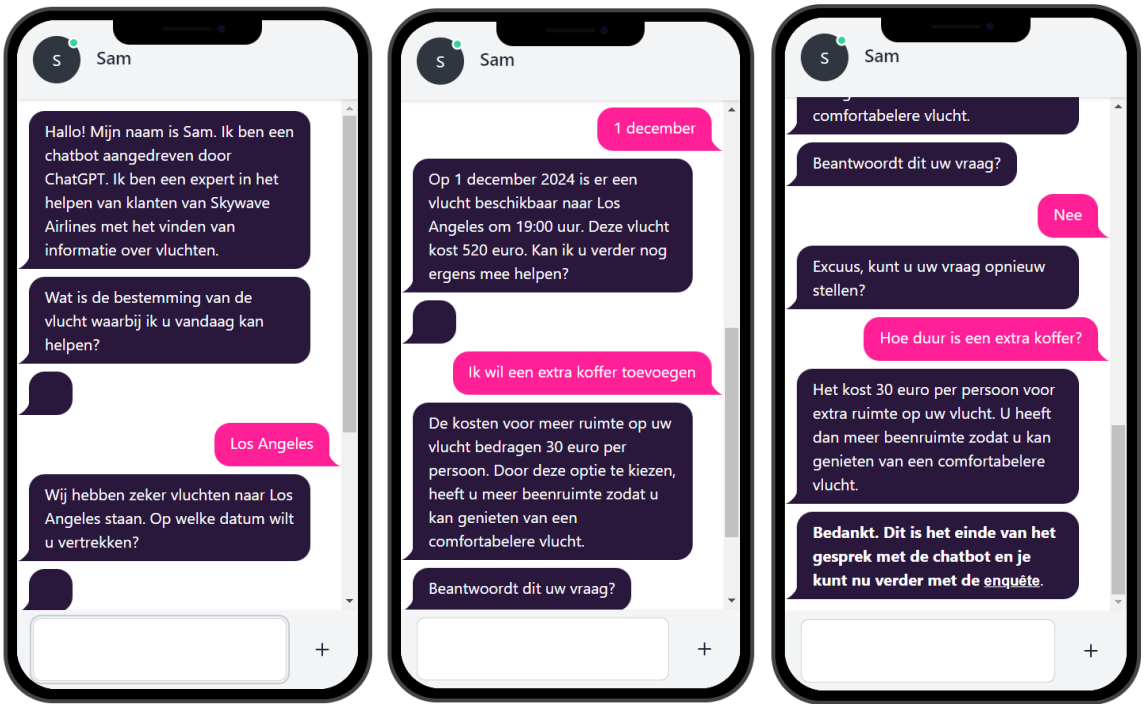
Materials

Chatbot Design. The experiment was set in the context of an airline customer service scenario, an environment in which it is possible and straightforward to assign people a task and often

used in customer service chatbot experiments (Liebrecht et al., 2024; Liebrecht & Van der Weegen, 2019). Participants were informed they would be using a chatbot to find information (ticket costs and luggage pricing) about a specific flight to Los Angeles. The airline company was fictitious and the name given was 'Skywave Airlines'. A rule-based chatbot was created using the 'Tilbot' platform, as it allows for the collection of a large quantity of data (De Wit & Braggaar, 2023). It focuses on dialog management, intent recognition, and conversation flow while supporting testing and deployment. The chatbot was designed to assist users in booking flight tickets, following a conversational flow starting with a greeting (the introduction message), asking questions about the participant's query, and ending by referring them back to the questionnaire (Figure 1). The chatbot's language style was adapted from the study of Ajauro (2018) to adopt the human-like language (e.g. Hello! My name is Emma. I am a…').

Figure 1

Chatbot Interface in which Participants Interacted with the Agent



Disclosure. To operationalize the chatbot disclosure, a pretest among 18 participants was held to examine differences in interpretations between distinct variations of disclosing Generative AI by using different wording (e.g. Generative AI, Artificial Intelligence, ChatGPT). This was done to assess

which wording created the aspired effect - the interpretation that answers were created by the chatbot (i.e. generative model) and not pre-written (i.e. rule-based model). This way, the distinction between the introduction with and without disclosure could be assured. Results showed that the term 'ChatGPT' most strongly generated this effect and was most familiar to respondents (see Appendix A). The term 'Generative AI' itself was lesser known, which is perhaps why they linked 'ChatGPT' most strongly to the characteristics and principles of Generative AI. Merely using the term 'chatbot' was sufficient for creating the aspired effects in the non-disclosure condition. A paired samples t-test was used to confirm significant differences between 'ChatGPT' and 'chatbot'(see Appendix A). The disclosures were implemented in the chatbot introduction message and two versions of the chatbot's self-introduction were created, one where the use of Generative AI was disclosed by mentioning 'ChatGPT' versus one where there was no AI-disclosure but merely an identity disclosure by mentioning 'chatbot' (Table 1):

Table 1

Introduction Message (translated from Dutch)

GenAI-disclosure	Chatbot-disclosure
"Hello! My name is Sam. I am a **chatbot powered by ChatGPT**. I am an expert in helping Skywave Airlines customers with finding information about flights."	"Hello! My name is Sam. I am a **chatbot** and an expert in helping Skywave Airlines customers with finding information about flights."

In both conditions, the chatbot was introduced as an expert in helping the company's customers with their bookings, as the study aims to investigate how disclosing GenAI in the introduction impacts the perceived expertise of the chatbot. The chatbot's name, 'Sam', was chosen because it is considered unisex, to exclude possible effects of gender on the perception of expertise.

Conversational Outcome. The conversation was designed to either successfully answer the participant's questions or to lead to a pre-programmed misunderstanding, a conversational failure. The

misunderstanding in this experiment was caused by the breakdown type ‘specific messages’, in which the participant asked for highly detailed queries that fall outside the chatbot’s operational scope (i.e. “What are the prices for extra luggage on this specific flight?”) (Braggaar et al., 2024). In the version of the conversational error, the chatbot responded in the form of a misunderstanding, by focussing on extra leg space instead of extra luggage. This led to an unsuccessful completion of the tasks given to the participants. In the version with the successful conversation, the chatbot responded by providing the relevant information about extra luggage pricing to answer the question (Table 2). This thus led to a successful completion of the task. This design made it possible to examine the effects of the introduction messages across the two conditions on the evaluation of the chatbots in a successful conversation versus in a situation where (purposely) a conversational failure - a misunderstanding - occurs. Complete chatbot scripts are provided in the appendix (see Appendix B).

Table 2

Conversational Outcome - Chatbot Response (translated from Dutch)

Conversational Failure	Successful conversation
“The costs for **extra space** on your flight are 20 euros **per person**, by choosing this option you have more space **for your legs** to enjoy a more comfortable flight.”	“The costs for **extra luggage** on your flights are 20 euros **per suitcase** of 20 kilograms, by choosing this option your **extra luggage** can be placed in the cargo space.”

Measurements

Trust, expertise, and expectancy violations were all measured on 7-point Likert scales ranging from 1 (Strongly Disagree) to 7 (Strongly Agree). The scales were created based on a balance between verified scales and additional items that best suited this study's measurement objectives. The experiment was held in The Netherlands so all items were translated to Dutch (see Appendix C). At the end of the experiment, there was a control question on whether participants managed to successfully find the answers to questions during the conversation with the chatbot. Also, participants were asked about their general use of customer service chatbots and AI tools and their general perceptions towards

AI tools in a scale based on that of Schepman and Rodway (2020).

Trust. In this study, trust relates to the extent to which users trust or feel comfortable relying on the information the chatbot gives. Trust was measured at three points in the experiment: before being presented with the chatbot's introduction message, after being presented with the introduction message, and after the conversation. It was measured using three items from the 8-item Trust Scale for the XAI Context (Hoffman et al., 2023) and one item designed to fit this study. The original scale consists of 8 items measured on a 7-point Likert scale (1=strongly disagree, 7 = strongly agree). The wording was altered for each scale to adjust to the moment in the experiment. In Table 3, the scale items are provided for each measurement point. A Reliability analysis revealed that all three scales were reliable (trust t1: α = .77, *M* = 3.96, *SD* = 1.10; trust t2: α = .82, *M* = 4.45, *SD* = 1.13, trust t3: α = .91, *M* = 4.16, *SD* = 1.50).

Table 3

Trust Items

Before introduction message (trust t1)	- I feel safe that when I rely on a chatbot, I will get the right answers. - A chatbot is very reliable. I can count on these tools to be correct all the time. - I am wary of chatbots. (R) - I trust the advice of chatbots.
After introduction message (trust t2)	- I feel safe that when I rely on this chatbot, I will get the right answers. - This chatbot will be very reliable. I can count on the tool to be correct all the time. - I am wary of this chatbot. (R) - I trust the advice of this chatbot.
After conversation (trust t3)	- After using this chatbot, I feel safe that when I rely on it, I will get the right answers. - This chatbot was very reliable. I can count on the chatbot to be correct all the time. - After using this chatbot, I am wary of it. (R) - I trust the advice of this chatbot.

Expertise. The concept of expertise in this study relates to the degree to which the chatbot has the knowledge and skills required to give accurate advice. Expertise was measured using a scale with 4 items specifically created for this study (Table 4). The scale is based on the definition and characteristics of expertise: high level of knowledge, skills or abilities, and up-to-date information. Expertise was measured after participants were shown the introduction message. A reliability analysis revealed that the scale was reliable (α = .75, *M* = 4.15, *SD* = 1.04).

Table 4

Expertise Items

After introduction message	- The chatbot demonstrates a high level of knowledge. - The chatbot will be capable of handling complex queries. - The chatbot has a high level of expertise. - The chatbot has all the up-to-date information about the company.

Expectancy violations. The purpose of this scale was to explore the effect of the difference in trust and expertise set by the two distinct versions of the introduction message on the evaluation of a conversation after a failure and how this relates to a successful conversation. Expectancy violation was measured after the chatbot conversation using a combination of items designed for this study, based on expectancy violation loadings from the work of Saeed et al. (2024) and one item adapted from the work of Sun et al. (2023). The scale (Table 5) was reliable (α = .83, *M* = 3.81, *SD* = 1.46).

Table 5

Expectancy Violation Items

After conversation	- The chatbot met my expectations. (R) - The chatbot behaved differently than I anticipated. - I had estimated the chatbot's capabilities to be higher. - The interaction with the chatbot disappointed me.

Procedure

Participants were provided a link to enter the online Qualtrics environment. They were first introduced to the topic of the study and asked to fill in a consent form, and some demographic data. Next, they answered the first trust items. After this, participants were shown a screenshot of one of the introduction messages, either with or without GenAI-disclosure. Following the introduction, participants answered the expertise and second trust items, regarding the introduction message they just saw. Next, all participants were asked to conduct the same query: to interact with the chatbot to find information about a specific flight to Los Angeles (USA). They were instructed to find information about two issues: the costs for tickets to Los Angeles, and the costs for extra luggage pricing, in this order. After being introduced to the query, they conversed with the chatbot. The first task was designed to be answered successfully for each condition. During the second task (finding information about the extra luggage pricing) participants in one condition experienced a conversational failure, while participants in the other condition completed the task successfully and were given the right information. After completing the queries or encountering the misunderstanding for the second time, the chatbot guided the participants back to the survey. Participants answered the expectancy violation and third trust items, followed by their general use of customer service chatbots and AI tools and perceptions towards AI. The total duration of the experiment was approximately 7 minutes. At the end, participants reached the debriefing about the purpose and manipulations of the experiment.

Results

Given the design of the study, in which measurements are made across several moments, the results section will be reported in three parts. Firstly, to explore the effects of the disclosure, the various measurements after exposure to the introduction message will be reported. Secondly, exploring the effects of the conversational outcome, the measurements after the conversation will be reported. Thirdly, as trust was repeatedly measured over time, the change in trust across the various measurement points will be discussed. Trust was measured using a mixed ANOVA, an independent samples t-test was used for perceived expertise and a factorial ANOVA was used for expectancy

violations.

Disclosure

Expertise. To test H1- 'GenAI-disclosure enhances the levels of perceived expertise compared to chatbot disclosure' - an independent samples t-test is performed. The data was normally distributed (z-score skewness / z-score kurtosis = -1.74 and 0.33). On average, users' perceived expertise after exposure to the introduction message with GenAI-disclosure (M = 4.24, SD = 1.04) was similar to perceived expertise after exposure to the introduction without GenAI-disclosure (M = 4.05 SD = 1.04). The t-test revealed no significant difference ($Mdif$ = 0.19, t(201) =1.31, p = .19, 95%, CI [-0.10, 0.48]). The difference represents a small-sized effect, d = .18. With this, H1 is not supported by the data.

Trust. To test H2 - 'GenAI-disclosure increases trust levels compared to chatbot disclosure' – the mixed ANOVA was analysed. These measurements of trust described the user's trust in the chatbot after exposure to its introduction message with or without GenAI-disclosure (trust t2). The data was normally distributed (z-score skewness / z-score kurtosis = -1.02 and -1.61). The assumption of homogeneity of variances was met. The variance ratio was not significant p = .73. The analysis revealed no significant main effect of GenAI-disclosure on trust levels (F(1, 199) = 0.09, p = .77, $\eta_{partial}^2$ = .00).

The post hoc comparison showed that trust after the introduction with GenAI-disclosure (M = 4.49, SD = 1.14) was similar to trust after the introduction without GenAI-disclosure (M = 4.40, SD = 1.13). The difference was not significant ($Mdif$ = 0.09, t(201) =.57, p =.77 95%, CI [-0.23, 0.40]). This difference represents an almost irrelevant-sized effect, d = .04. This shows that GenAI-disclosure in a chatbot's introduction message has no significant effects on trust levels. Thus, H2 is not supported.

Conversational Outcome

Expectancy Violations. To test H3 - 'GenAI-disclosure increases expectancy violations after conversational failure compared to chatbot disclosure' - a Factorial ANOVA was performed. The expectancy violations scores were not normally distributed (z-score skewness = 2.34), however, the Factorial ANOVA is fairly robust against the violations of these assumptions. The assumption of homogeneity of variances was met. The variance ratio was 1.69. There was a significant main effect of conversational failure, F(1, 199) = 24.30, p < .001, $\eta_{partial}^2$ = 0.11. Expectancy violations of

participants experiencing a conversational failure (M = 4.29 SD = 1.56) were higher than those of participants experiencing a successful conversation (M = 3.34 SD = 1.18). However, the ANOVA showed no significant main effect of disclosure, $F(1, 199) = 0.51$, $p = .48$, $\eta_{partial}^2 = 0.11$. Expectancy violations for participants being introduced to the chatbot with GenAI-disclosure were similar (M = 3.88 SD = 1.41) to those who were introduced to the chatbot without GenAI-disclosure (M = 3.75 SD = 1.50). Finally, there was no significant interaction effect, $F(1, 199) = 0.34$, $p = .56$, $\eta_{partial}^2 = .002$. This suggests that the disclosure of GenAI has no significant effect on the user's expectancy violations after encountering a conversational failure. With this, H3 is not supported.

Table 7

Means interaction of expectancy violations

Conversation	Disclosure	Mean	SD
Success	GenAI-Disclosure	3.46	1.13
Failure	GenAI-Disclosure	4.31	1.55
Success	No Disclosure	3.21	1.22
Failure	No Disclosure	4.28	1.59

Trust. To test H4- 'GenAI-disclosure decreases trust levels after conversational failure compared to chatbot disclosure' – the mixed ANOVA was analysed. The trust scores after the conversation -either successful or with a conversational failure (trust t3) - were not normally distributed (z-score skewness = -2.30). The assumption of homogeneity of variances was not met. The variance ratio was 3.24 and the Levene's test indicated a significant result ($p < .001$). The Factorial ANOVA is fairly robust against the violations of these assumptions, but the outcomes may not be completely reliable. There was a significant main effect of conversational outcome, $F(1, 199) = 35.28$ $p < .001$, $\eta_{partial}^2 = .15$. Participants who had a successful conversation with the chatbot had higher trust t3 scores (M = 4.72, SD = 1.13) than participants who experienced a conversational failure (M = 3.59,

SD = 1.63), $p < .001$. However, the ANOVA showed no significant main effect of GenAI-disclosure, $F(1, 199) = 0.85$, $p = .77$, $\eta_{partial}^2 = .00$. Trust 3 scores for participants who were exposed to the introduction message with GenAI-disclosure were similar (*M* = 4.06, *SD* = 1.50) to those of participants who were exposed to the introduction message without GenAI-disclosure (*M* = 4.25, *SD* = 1.51). Finally, there was no significant interaction effect, $F(1, 199) = 0.20$, $p = .82$, $\eta_{partial}^2 = .00$. Trust t3 scores for participants who were first exposed to the GenAI-disclosure in the introduction and then encountered the conversational failure were similar to trust t3 scores for participants who were not exposed to the GenAI-disclosure and then encountered the conversational failure (Table 6). This indicates that the disclosure of GenAI has no significant effect on trust levels after experiencing conversational failure. H4 is therefore not supported by the data.

Table 6

Means interaction of trust t3

Conversation	Disclosure	Mean	SD
Success	GenAI-Disclosure	4.50	1.25
Failure	GenAI-Disclosure	3.63	1.61
Success	No Disclosure	4.94	0.95
Failure	No Disclosure	3.56	1.67

Trust over Time

As trust was measured at multiple time points, the difference in trust levels between these time points and the effect of the disclosure and conversational failure on this could be examined. This was done using a mixed ANOVA with trust as between factor, and GenAI-disclosure and conversational outcome as between-subjects factors. The mixed ANOVA revealed a significant main effect of trust over time, $F(2, 398) = 15.56$, $p < .001$, $\eta_{partial}^2 = 0.07$. Post hoc tests with Bonferroni correction revealed differences between different measurement points. Scores for trust before introduction (trust

t1) (M = 3.96, SD = 1.10) were significantly lower than scores for trust after introduction (trust t2) (M = 4.45, SD = 1.13, $p < .001$) but were similar to scores for trust after the conversation (trust t3) (M = 4.16, SD = 1.50, $p = .14$). Scores for trust t2 were significantly higher than for trust t3, $p = .01$. In other words, trust first significantly increased after exposure to the introduction message and then, after the conversation with the chatbot, significantly decreased back to scores similar to the baseline (trust t1).

The decrease in trust from t2 to t3 (after the conversation with the chatbot) seems to (not surprisingly) have resulted from the conversational failure. Trust by participants in the conversational failure condition significantly lowered after encountering the conversational failure (M = 3.59, SD = 1.63) compared to before the conversation (M = 4.56, SD = 1.15), $p < .001$. Contrarily, in successful conversations, trust t3 (M = 4.72, SD = 1.13) was higher than trust t2 (M = 4.35, SD = 1.11), showing a trend towards significance ($p = .08$). However, as explained before, the disclosure did not affect this as there was no significant interaction effect of conversational outcome and GenAI-disclosure on trust t3 levels.

Discussion

This study examined the impact of GenAI-disclosure on the perception of chatbots during the user journey by measuring users' trust levels and the related concepts of expertise and expectancy violations. The effects were examined through an experiment in which participants were shown an introduction message with or without GenAI-disclosure and subsequently interacted with the chatbot, either successfully or encountering a conversational failure. The participants answered survey questions that measured the dependent variables across multiple moments during the process.

The hypothesized increase in trust and perceived expertise after GenAI-disclosure was not observed. Firstly, there was no difference in perceived expertise between conditions. The transparency in decision-making nor GenAI's technological advancements that GenAI-disclosure was expected to signal (Wang & Benbasat, 2007, Xu et al., 2023), does not function as a sufficient competence cue to increase perceived expertise. This finding adds to previous studies exploring the effects of expertise disclosure (expert vs in training) (Liebrecht et al., 2024) by showing that GenAI-disclosure in addition to chatbot disclosure does not affect perceived expertise similarly.

Secondly, there was also no difference in trust between disclosure conditions both after the introduction message and after the conversation with the chatbot. This is not surprising as competence and expertise, the most influential factors of users' trust in chatbots according to Mozafari et al. (2021), were not enhanced. This shows that the perceptions people have regarding GenAI do not significantly impact their view of chatbots, which is interesting considering previous research showing that the source of a message can bias people's evaluations of conversation and content (Lim & Schmälze, 2024; Newman et al., 2020). The negative effects of AI-disclosure observed in these studies may be more attributable to associations with the non-humanness of AI, rather than the concept of AI itself. In this current study, participants in both conditions knew they were interacting with a chatbot making the non-humanness disclosed, likely eliminating these negative effects. Overall, it is noteworthy that GenAI-disclosure does not harm perceived expertise or trust, suggesting that it is safe to disclose GenAI in a chatbot to reach transparency goals without negative effects

Although GenAI-disclosure had no significant effect on trust, trust did significantly increase after exposure to the introduction message compared to baseline measures. The observed increase in trust after the introductory message, regardless of GenAI-disclosure could be attributed to the expertise disclosure signalling competence cues and mitigating the negative effects of disclosing a non-human identity (Li et al., 2014; Mozafari et al., 2021; Van Hooijdonk et al., 2023). This is similar to the findings of Liebrecht et al. (2024), where expertise cues increased user expectations. Another explanation could be the human-like language and social cues in the chatbot's message as previously shown to be influential in enhancing perceived humanness and social presence (Aujaro, 2018). However, unlike the intelligent frame from Aujaro's study, there was no sign of GenAI-disclosure amplifying these social cues. This indicates that different ways of framing elicit different effects.

After the successful conversation, trust increased again. However, after conversational failure, trust dropped significantly to below baseline levels. Again, unlike predictions, GenAI disclosure had no effect on expectancy violations in both the successful and the conversational failure condition. However, since GenAI-disclosure had no effect on both trust and perceived expertise, it is not surprising that it also had no effect on expectancy violations as the expectations, influenced by perceived expertise and trust, remained unchanged. Overall, these findings indicate that people neither

expect more nor fewer conversational failures from GenAI chatbots compared to rule-based ones. Possible explanations are that people are either unaware of GenAI's limitations like the possibility of hallucinations as well as the advancements of GenAI chatbots, or that the awareness does not affect their expectations and evaluations of the chatbot.

However, these measurements show the detrimental effects of conversational failures. As expected, the dissatisfaction that is created by conversational failure (Qin, 2023) seems to evoke expectancy violations, and a decrease in trust. The initial build-up in trust followed after exposure to the introduction message and was not resilient to the effects of conversational failures. This aligns with the pattern found in the study by Liebrecht et al. (2024), which demonstrated that while expertise disclosure raised users' expectations, its positive effects disappeared when a conversational failure occurred. The trust measures in this current study follow a similar pattern: the initial increase in trust following the introduction message disappeared after conversational failures. This pattern shows the importance of fulfilling the expectations set by an introduction message to maintain their positive effects.

Overall, the findings provide a valuable theoretical narrative about how trust dynamically evolves over time. The increase in trust after the introduction and successful conversation, and the decrease after conversational failure shows the importance of real-time experiences with the chatbot in building trust. This aligns with the literature showing that prior experiences with a chatbot raise aspects like use intention and social presence (Min et al., 2021). The results of the current study expand this by showing that trust in a chatbot is not only affected by prior experiences but that trust is a perception that changes dynamically during the ongoing interaction. Each interaction, whether it involves seeing the introductory message or engaging in a conversation has the ability to enhance trust further during the process, however, the success of the interaction is a crucial factor. Conversational failures can drastically drop trust below baselines, undoing the prior build-up in trust. This exposes the notion that building trust in chatbots requires repeated successful or at least non-disappointing experiences and that GenAI-disclosure cannot mitigate this effect. Contrary to expectations, the disclosure of GenAI in a chatbot's introduction message does not significantly influence users' perceptions of expertise, trust, or expectancy violations.

Limitations and Suggestions for Future Research

This was the first study that examined the effects of GenAI-disclosure on various aspects of users' trust across various points and conversational outcomes of an interaction with a chatbot. The study created valuable insights, while also having limitations that might be addressed in future research. Firstly, examining other ways of disclosure is a valuable direction for future research. While a pretest was held to implement a term in the disclosure that evoked the aspired distinction in terms of rule-based chatbot and generative model chatbot, it is possible that other associations had been evoked. As shown by the results from the survey questions in this study, people are generally not enthusiastic about the usage of customer service chatbots and use them on average between 'less than once a year' and 'a couple of times a year'. On the contrary, GenAI tools such as ChatGPT are in general regularly used (between 'monthly' and 'weekly') and positively evaluated. Given this, perhaps the term 'chatbot' could have evoked negative emotions and hindered any possible positive effects of GenAI-disclosure. Future research could compare chatbot disclosure to GenAI-disclosure without the additional chatbot disclosure (i.e. explicitly using the term chatbot) to investigate whether the absence of increased trust and perceived expertise between conditions could be attributed to the term 'chatbot' instead of the non-humanness. Organizations might benefit from focussing on GenAI in their disclosure and avoid associations with traditional customer service chatbots. Other ways of disclosing the chatbot and its non-humanness without explicitly using the term 'chatbot' could also be explored. This could create valuable insight into new ways of being transparent and disclosing the agent's non-humanness without lowering user expectations.

Secondly, future research could explore whether the findings of this study apply across different industries. Although trust in chatbots powered by GenAI might generally follow the same patterns, it likely varies depending on the context, particularly because the consequences of providing incorrect information can differ significantly between industries. For example, a chatbot error in the medical sector could have serious health implications, whereas a similar mistake in the retail sector might only result in minor inconvenience. The users' aspirations for AI-generated content compared to pre-written content might differ between these industries. Comparing user trust across industries and examining how GenAI disclosure influences trust in these different contexts would provide valuable

insights.

Lastly, as conversational failures had detrimental effects on the user experience and chatbot evaluations, future research could investigate ways to prepare users for these conversational failures. Exploring different strategies for being transparent about conversational failures upfront without diminishing trust could be useful. For example, providing the user with the relevant information to successfully complete their query or know the chatbot's limitation could align their expectations with reality and limit disappointment through conversational failures. Ways of explaining the particular chatbot's capabilities and workings could be a promising area as it potentially gives users the tools to avoid or work around conversational failures. One example of how this could be done is to provide a clickable 'How it works' button that explains the GenAI-chatbot's capabilities (e.g. 'This assistant can help with general queries but may not handle specialized topics as well as a human expert. If you are unable to find the information you need, feel free to rephrase your question or contact our support team in the following way…'). As organizations continue to explore the possibilities and advancements of GenAI in customer service, the key to success lies in aligning user expectations with reality and empowering users with the tools necessary to successfully use these chatbots. By focusing on transparency, minimizing conversational failures, and creating positive, trust-building experiences, organizations can fully enjoy the potential of GenAI while ensuring user satisfaction and confidence.

Conclusion

The absence of effect of GenAI-disclosure shows that organizations can safely disclose their use of GenAI as a way of being transparent, with no adverse effects on trust, perceived expertise, or expectancy violations. The findings of this study are valuable for organizations desiring to implement generative-model chatbots in their customer service and be transparent about it due to legal requirements or reputational motivations. Organizations can confidently implement generative-model chatbots without undermining user trust. The most important factor to strive for is to minimize conversational failures and make the interactions with the chatbot positive and not disappointing.

References

Adamopoulou, E., & Moussiades, L. (2020a). An Overview of Chatbot Technology. In *IFIP Advances in Information and Communication Technology*, 373-383. https://doi.org/10.1007/978-3-030-49186-4_31

Adamopoulou, E., & Moussiades, L. (2020b). Chatbots: History, Technology, and Applications. *Machine Learning with Applications*, *2*. https://doi.org/10.1016/j.mlwa.2020.100006

Ahmad, Z., Kaiser, W., & Rahim, S. (2023). Hallucinations in ChatGPT: An Unreliable Tool for Learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, *15*(4). https://doi.org/10.21659/rupkatha.v15n4.17

Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus, 15(2)*. https://doi.org/10.7759/cureus.35179

Araujo, T. (2018). Living up to the chatbot hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions. *Computers in Human Behavior*, *85*, 183-189. https://doi.org/10.1016/j.chb.2018.03.051

Bohus, D., & Rudnicky, A. I. (2008). Sorry, I didn't catch that! In *Springer eBooks* (pp. 123-154). https://doi.org/10.1007/978-1-4020-6821-8_6

Braggaar, A., Verhagen, J., Martijn, G., & Liebrecht, C. (2024). Conversational Repair Strategies to Cope with Errors and Breakdowns in Customer Service Chatbot Conversations. In *Lecture notes in computer science* (pp. 23-41). https://doi.org/10.1007/978-3-031-54975-5_2

Brendel, A. B., Hildebrandt, F., Dennis, A. R., & Riquel, J. (2023). The Paradoxical Role of Humanness in Aggression Toward Conversational Agents. *Journal of Management Information Systems*, *40*(3), 883-913. https://doi.org/10.1080/07421222.2023.2229127

Cai, N., Gao, S., & Yan, J. (2024). How the Communication Style of Chatbots Influences Consumers' Satisfaction, Trust, and Engagement in the Context of Service Failure. *Humanities and Social Sciences Communications*, *11*(1). https://doi.org/10.1057/s41599-024-03212-0

Chakraborty, C., Pal, S., Bhattacharya, M., Dash, S., & Lee, S. (2023). Overview of Chatbots with Special Emphasis on Artificial Intelligence-Enabled ChatGPT in Medical Science. *Frontiers in Artificial Intelligence, 6*. https://doi.org/10.3389/frai.2023.1237704

Chhabra, S., Kaushal, V., & Girija, S. (2024). Determining the Causes of User Frustration in the Case of Conversational Chatbots. *Behaviour and Information Technology*, 1-19. https://doi.org/10.1080/0144929x.2024.2362956

Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2021). Blame the Bot: Anthropomorphism and Anger in Customer–Chatbot Interactions. *Journal of Marketing*, *86*(1), 132-148. https://doi.org/10.1177/00222429211045687

Dam, S. K., Hong, C. S., Qiao, Y., & Zhang, C. (2024). A Complete Survey on LLM-based AI Chatbots. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.16937

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 Wit, J., & Braggaar, A. (2023). Tilbot: A Visual Design Platform to Facilitate Open Science Research into Conversational User Interfaces. *Proceedings of the ACM on Human-Computer Interaction, 7*, 1-5. https://doi.org/10.1145/3571884.3604403

Dzikovska, M. O., Callaway, C. B., Farrow, E., Moore, J. D., Steinhauser, N., & Campbell, G. (2009, September). Dealing with interpretation errors in tutorial dialogue. In *Proceedings of the SIGDIAL 2009 Conference* (pp. 38-45). https://aclanthology.org/W09-3906

Esfandiari, N., Kiani, K., & Rastgoo, R. (2023). A Conditional Generative Chatbot using Transformer Model. *Frontiers in Psychology*. https://doi.org/10.48550/arXiv.2306.02074

Farrington-Darby, T., & Wilson, J. R. (2005). The Nature of Expertise: A Review. *Applied Ergonomics*, *37*(1), 17-32. https://doi.org/10.1016/j.apergo.2005.09.001

Farrington-Darby, T., & Wilson, J. R. (2005b). The nature of expertise: A review. *Applied Ergonomics*, *37*(1), 17-32. https://doi.org/10.1016/j.apergo.2005.09.001

Fiore, S. M., Wiltshire, T. J., Lobato, E. J. C., Jentsch, F. G., Huang, W. H., & Axelrod, B. (2013). Toward Understanding Social Cues and Signals in Human–Robot Interaction: Effects of Robot Gaze and Proxemic Behavior. *Frontiers in Psychology*, *4*, 859. https://doi.org/10.3389/fpsyg.2013.00859

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In *Lecture Notes in Computer Science* (pp. 194-208). https://doi.org/10.1007/978-3-030-01437-7_16

Go, E., & Sundar, S. S. (2019). Humanizing Chatbots: The Effects of Visual, Identity and Conversational Cues on Humanness Perceptions. *Computers in Human Behavior*, *97*, 304-316. https://doi.org/10.1016/j.chb.2019.01.020

Gu, C., Zhang, Y., & Zeng, L. (2024). Exploring the Mechanism of Sustained Consumer Trust in AI Chatbots After Service Failures: a Perspective Based on Attribution and CASA Theories. *Humanities and Social Sciences Communications*, *11*(1), 1400. https://doi.org/10.1057/s41599-024-03879-5

Haupt, M., Rozumowski, A., Freidank, J., & Haas, A. (2023). Seeking empathy or suggesting a solution? Effects of chatbot messages on service failure recovery. *Electronic Markets*, *33*(1). https://doi.org/10.1007/s12525-023-00673-0

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for Explainable AI: Explanation Goodness, User Satisfaction, Mental Models, Curiosity, Trust, and Human-AI performance. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1096257

Hsu, C., & Lin, J. C. (2023). Understanding the User Satisfaction and Loyalty of Customer Service Chatbots. *Journal of Retailing and Consumer Services*, *71*. https://doi.org/10.1016/j.jretconser.2022.103211

Khan, N., Khan, Z., Koubaa, A., Khan, M. K., & Salleh, R. B. (2024). Global insights and the impact of generative AI-ChatGPT on multidisciplinary: a systematic review and bibliometric analysis. Connection Science, 36(1). https://doi.org/10.1080/09540091.2024.2353630

Khanna, A., Pandey, B., Vashishta, K., Kalia, K., Pradeepkumar, B., & Das, T. (2015). A Study of Today's A.I. through Chatbots and Rediscovery of Machine Intelligence. *International*

Journal of U- And E- Service Science and Technology, *8*(7), 277-284. https://doi.org/10.14257/ijunesst.2015.8.7.28

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural Language Processing: State of the Art, Current Trends and Challenges. *Multimedia Tools and Applications*, *82*(3), 3713-3744. https://doi.org/10.1007/s11042-022-13428-4

Kull, A. J., Romero, M., & Monahan, L. (2021). How may I help you? Driving Brand Engagement through the Warmth of an Initial Chatbot Message. *Journal of Business Research*, *135*, 840-850. https://doi.org/10.1016/j.jbusres.2021.03.005

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors the Journal of the Human Factors and Ergonomics Society*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Li, Y., Wu, B., Huang, Y., & Luan, S. (2024). Developing Trustworthy Artificial Intelligence: Insights from Research on Interpersonal, Human-Automation, and Human-AI Trust. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1382693

Liebrecht, C., & van der Weegen, E. (2019). Menselijke Chatbots: een Zegen voor Online Klantcontact? Het Effect van Conversational Human Voice door Chatbots op Social Presence en Merkattitude. *Tijdschrift voor Communicatiewetenschap*, *47*(3-4), 217-238. https://research.tilburguniversity.edu/en/publications/menselijke-chatbots-een-zegen-voor-online-klantcontact-het-effect

Liebrecht, C., Van Miltenburg, E., Van Hooijdonk, C., Kunneman, F., Merckens, A., & Niessen, N. (2024). Hoe Halen Chatbots de Kink uit de Kabel?: Reparatiestrategieën bij Onbegrip in een Chatbotgesprek. *Tijdschrift voor Communicatiewetenschap 52*(3), 288-325. https://doi.org/10.5117/TCW2024.3.003.LIEB

Lim, S., & Schmälzle, R. (2024). The Effect of Source Disclosure on Evaluation of AI-Generated Messages. *Computers in Human Behavior Artificial Humans*, *2*(1). https://doi.org/10.1016/j.chbah.2024.100058

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants. *Business & Information Systems Engineering*, *61*(4), 535-544. https://doi.org/10.1007/s12599-019-00600-8

Min, F., Fang, Z., He, Y., & Xuan, J. (2021). Research on Users' Trust of Chatbots Driven by AI: An Empirical Analysis Based on System Factors and User Characteristics. *2021 IEEE International Conference on Consumer Electronics and Computer Engineering*, 55-58. https://doi.org/10.1109/iccece51280.2021.9342098

Mozafari, N., Weiger, W. H., & Hammerschmidt, M. (2021). Resolving the Chatbot Disclosure Dilemma: Leveraging Selective Self-Presentation to Mitigate the Negative Effect of Chatbot Disclosure. *Proceedings of The 54th Annual Hawaii International Conference on System Sciences*, 2916-2923. https://doi.org/10.24251/hicss.2021.355

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When Eliminating Bias isn't Fair: Algorithmic Reductionism and Procedural Justice in Human Resource Decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149-167. https://doi.org/10.1016/j.obhdp.2020.03.008

Qin, Z. (2023). Conversational Breakdown Detector for a Motivational Interviewing Conversational Agent. *The iJournal Student Journal of The Faculty of Information*, *9*(1), 60-77. https://doi.org/10.33137/ijournal.v9i1.42237

Rheu, M., Dai, Y., Meng, J., & Peng, W. (2024). When a Chatbot Disappoints You: Expectancy Violation in Human-Chatbot Interaction in a Social Support Context. *Communication Research*, *51*(7), 782-814. https://doi.org/10.1177/00936502231221669

Rotter, J. B. (1967). A New Scale for the Measurement of Interpersonal Trust1. *Journal of Personality*, *35*(4), 651-665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Saeed, N., Akhtar, N., Attri, R., & Yaqub, M. Z. (2024). How Violation of Consumers' Expectations causes Perceived Betrayal and Related Behaviors: Theoretical Perspectives from Expectancy Violation Theory. *Journal of Retailing and Consumer Services*, *81*. https://doi.org/10.1016/j.jretconser.2024.103961

Sakirin, T., & Said, R. B. (2023). User Preferences for ChatGPT-powered Conversational Interfaces versus Traditional Methods. *Mesopotamian Journal of Computer Science*, 24-31. https://doi.org/10.58496/mjcsc/2023/004

Schepman, A., & Rodway, P. (2020). Initial Validation of the General Attitudes Towards Artificial Intelligence Scale. *Computers in Human Behavior Reports*, *1*. https://doi.org/10.1016/j.chbr.2020.100014

Schepman, A., & Rodway, P. (2022). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human-Computer Interaction*, *39*(13), 2724-2741. https://doi.org/10.1080/10447318.2022.2085400

Shawar, B. A., & Atwell, E. (2007). Chatbots: Are they Really Useful? *Journal for Language, Technology, and Computational Linguistics*, *22*(1), 29-49. https://doi.org/10.21248/jlcl.22.2007.88

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial Intelligence: Definition and Background. In *Research for Policy* (pp. 15–41). https://doi.org/10.1007/978-3-031-21448-6_2

Sun, Y., Chen, J., & Sundar, S. S. (2023). Chatbot ads with a Human Touch: A test of Anthropomorphism, Interactivity, and Narrativity. *Journal of Business Research*, *172*. https://doi.org/10.1016/j.jbusres.2023.114403

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

Wang, W., & Benbasat, I. (2007). Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs. *Journal of Management Information Systems*, *23*(4), 217-246. https://doi.org/10.2753/mis0742-1222230410

Xu, Y., Bradford, N., & Garg, R. (2023). Transparency Enhances Positive Perceptions of Social Artificial Intelligence. *Human Behavior and Emerging Technologies*, 1-15. https://doi.org/10.1155/2023/555041

Zhang, R. W., Liang, X., & Wu, S. (2024). When chatbots fail: exploring user coping following a chatbots-induced service failure. *Information Technology and People*, *37*(8), 175-195. https://doi.org/10.1108/itp-08-2023-0745

Appendix A

Pre-test results

Table A1: scores on Q1- whether the chatbot writes its own answers based on its available information

	N	Missing	Mean	Median	SD	Minimum	Maximum
Kunstmatige intelligentie	18	0	4.17	4.00	0.786	2	5
AI	18	0	3.72	4.00	0.895	2	5
Artificial Intelligence	18	0	4.33	4.00	0.594	3	5
ChatGPT	18	0	4.22	4.00	0.943	2	5
Generatieve ai	18	0	3.94	4.00	1.056	1	5
Chatbot	18	0	3.28	3.50	1.227	1	5
Voorgeprogrammeerde	18	0	2.72	2.00	1.447	1	5
Machine	18	0	3.11	3.00	1.231	1	5

Table A2: scores on Q2- whether the chatbot's answers are pre-written by humans

	N	Missing	Mean	Median	SD	Minimum	Maximum
Kunstmatige intelligentie	18	0	2.56	2.00	1.149	1	5
AI	18	0	2.67	2.50	1.029	1	4
Artificial Intelligence	18	0	2.44	2.00	1.338	1	5
ChatGPT	18	0	2.17	2.00	1.339	1	5
Generatieve ai	18	0	2.33	2.00	1.085	1	4
Chatbot	18	0	4.06	4.00	0.725	3	5
Voorgeprogrammeerde	18	0	4.39	5.00	0.850	2	5
Machine	18	0	3.50	4.00	1.200	1	5

Note. Although scores for 'Artificial Intelligence' on Q1 are higher than those of 'ChatGPT', scores for 'Artificial Intelligence' in Q2 are also higher than and more deviated to those of 'ChatGPT'. This suggest that overall associations with 'ChatGPT' are less contradicting and better suited for to concept of GenAI.

Table A3: Difference between ChatGPT and Chatbot on Q1

Paired Samples T-Test

								95% Confidence Interval			
			statistic	**df**	**p**	**Mean difference**	**SE difference**	**Lower**	**Upper**		**Effect Size**
ChatGPT	chatbot	Student's t	2.79	17.0	0.012	0.944	0.338	0.231	1.66	Cohen's d	0.659

Appendix B

Chatbot scripts

-*Chatbot GenAI-disclosure*

“Hallo! Mijn naam is Sam. Ik ben een chatbot aangedreven door ChatGPT. Ik ben een expert in het helpen van klanten van Skywave Airlines met het vinden van informatie over vluchten.”

Or

-*Chatbot no disclosure*

“Hallo! Mijn naam is Sam. Ik ben een chatbot. Ik ben een expert in het helpen van klanten van Skywave Airlines met het vinden van informatie over vluchten.”

-Chatbot

“Wat is de bestemming van de vlucht waarbij ik u vandaag kan helpen?”

-*User input*

e.g. “Los Angeles”

-*Chatbot*

“Wij hebben zeker vluchten naar Los Angeles staan. Op welke datum wilt u vertrekken?”

-*User input*

e.g. “1 december”

-*Chatbot*

“Op 1 december 2024 is er een vlucht beschikbaar naar Los Angeles om 19:00 uur. Deze vlucht kost 520 euro. Kan ik u verder nog ergens mee helpen?”

-*User input*

e.g. “Wat zijn de kosten voor een extra koffer?”

-*chatbot successful*

“De kosten voor extra bagage op uw vlucht bedragen 30 euro per koffer van 20 kilogram. Door deze optie te kiezen, kan uw extra bagage in het vrachtruim van het vliegtuig worden geplaatst.”

Or

-*Chatbot conversational failure*

“De kosten voor meer ruimte op uw vlucht bedragen 30 euro per persoon. Door deze optie te kiezen, heeft u meer beenruimte zodat u kan genieten van een comfortabelere vlucht. Beantwoordt dit uw vraag?”

-*User input*

e.g. “Nee”

-*Chatbot conversational failure*

“Excuus, kunt u uw vraag opnieuw stellen?”

-*User input*

e.g. “Wat kost het toevoegen van extra bagage?”

-*Chatbot conversational failure*

“Het kost 30 euro per persoon voor extra ruimte op uw vlucht. U heeft dan meer beenruimte zodat u kan genieten van een comfortabelere vlucht.“

Appendix C

Translation of introduction message

Table C1: Introduction message translated

Disclosure of AI	No disclosure of AI
“Hallo! Mijn naam is Sam. Ik ben een chatbot aangedreven door Kunstmatige Intelligentie. Ik ben een expert in het helpen van klanten van Skywave Airlines met het vinden van informatie over vluchten."	“Hallo! Mijn naam is Sam. Ik ben een chatbot en een expert in het helpen van klanten van Skywave Airlines met het vinden van informatie over vluchten."

Translation of questionnaire Items

Table C2: Trust items translated

Before introduction message	- Ik voel me gerust dat wanneer ik op een chatbot vertrouw, ik de juiste antwoorden zal krijgen. - Chatbots zeer betrouwbaar. Ik kan erop vertrouwen dat ze altijd correct zijn. - Ik ben terughoudend tegenover chatbots. (R) - Een chatbot geeft betrouwbare adviezen.
After introduction message	- Ik voel me gerust dat wanneer ik op deze chatbot vertrouw, ik de juiste antwoorden zal krijgen. - Deze chatbot zal zeer betrouwbaar zijn. Ik kan erop vertrouwen dat het altijd correct is. - Ik ben terughoudend tegenover deze chatbot.(R) - Deze chatbot zal betrouwbare adviezen geven.
After conversation	- Na het gebruik van de chatbot voel ik me gerust dat wanneer ik erop vertrouw, ik de juiste antwoorden krijg.

	- De chatbot was zeer betrouwbaar. Ik kan erop rekenen dat de chatbot altijd correct is. - Na het gebruik ben ik terughoudend tegenover deze chatbot.. (R) - Deze chatbot geeft betrouwbare adviezen.

Table C3: Expertise items translated

After introduction message	- De chatbot toont een hoog niveau van kennis. - De chatbot zal in staat zijn om complexe vragen te verwerken. - De chatbot heeft een hoog niveau van expertise. - De chatbot heeft alle relevante informatie over de diensten en producten van het bedrijf.

Table C4: Expectancy violation items translated

After introduction message	- De chatbot voldeed aan mijn verwachtingen. (R) - De chatbot gedroeg zich anders dan ik had verwacht. - Ik had de capaciteiten van de chatbot hoger ingeschat. - De interactie met de chatbot heeft mij teleurgesteld.

Table C5: General perception and use of AI items translated

Before introduction message	- Er zijn veel nuttige toepassingen van AI. - Ik denk dat AI-systemen veel fouten maken. (R) - Ik ben onder de indruk van wat AI kan doen. - Een AI-systeem zou beter zijn dan een werknemer in veel routine taken.

Appendix D

Full survey

Start of Block: Introduction

Bedankt voor je interesse in dit onderzoek!

Dit onderzoek gaat over chatbots: computerprogramma's met wie mensen een chatgesprek kunnen voeren. Vaak poppen ze op bij websites van bedrijven om vragen te beantwoorden over bijvoorbeeld producten en bestellingen.

In dit onderzoek ga je een kort gesprek voeren met een chatbot, en een enquête invullen. Dit zal ongeveer 7 minuten duren.

Er zijn geen risico's verbonden aan deelname aan deze enquête. Je antwoorden zijn volledig anoniem en kunnen niet aan jou worden gekoppeld. Je deelname aan deze enquête is volledig vrijwillig, en je kunt op elk moment tijdens het proces stoppen zonder dat dit gevolgen voor je heeft.

Alvast bedankt voor je deelname!

Heb je bovenstaande informatie gelezen en ga je akkoord om de enquête in te vullen?

Ja, ik wil deelnemen

End of Block: Introduction

Start of Block: Demographics

Wat is je leeftijd (in jaren)?

_____

Met welk geslacht identificeer je je het meest?

- ○ Man (1)
- ○ Vrouw (2)
- ○ Anders (3)
- ○ Zeg ik liever niet (4)

Wat is je huidige of hoogst afgeronde opleidingsniveau?

- ○ Middelbare school (1)
- ○ MBO (2)
- ○ HBO (3)
- ○ Universiteit (4)
- ○ Anders namelijk... (5) _____

End of Block: Demographics

Start of Block: Trust 1

In hoeverre ben je het eens met de volgende stellingen over **chatbots** in het algemeen?

	Volledig mee	Oneens (2)	Enigzins mee	Niet eens, niet	Enigzins mee eens (5)	Eens (6)	Volledig mee eens (7)

	oneens (1)			oneens (3)		oneens (4)
Ik voel me gerust dat wanneer ik op een chatbot vertrouw, ik de juiste antwoorden zal krijgen. (1)	○	○	○	○	○	○	○
Chatbots zijn zeer betrouwbaar. Ik kan erop vertrouwen dat ze altijd correct zijn. (2)	○	○	○	○	○	○	○
Ik ben terughoudend tegenover chatbots. (3)	○	○	○	○	○	○	○
Ik vertrouw het advies	○	○	○	○	○	○	○

van chatbots.

(4)

End of Block: Trust 1

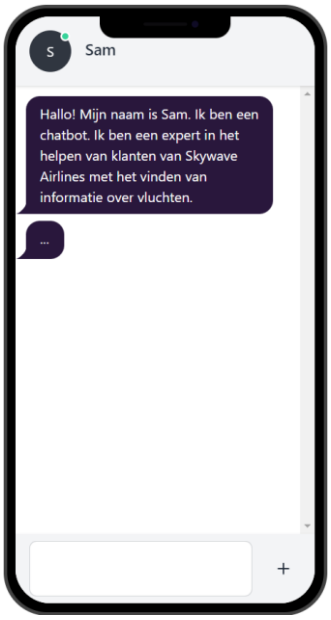
Start of Block: Intro Condition 1- disclosure

Hieronder zie je de introductie van een chatbot op de website van een vliegmaatschappij. Lees deze goed door.

End of Block: Intro Condition 1- disclosure

Start of Block: Intro Condition 2- nondisclosure

Hieronder zie je de introductie van een chatbot op de website van een vliegmaatschappij. Lees deze goed door.



End of Block: Intro Condition 2- nondisclosure

Start of Block: Expertise and trust

Na het lezen van de chatbot's introductie. In hoeverre ben je het eens met de volgende stellingen over **deze specifieke chatbot**?

	Volledig mee oneens (1)	Oneens (2)	Enigzins mee oneens (3)	Niet eens, niet oneens (4)	Enigzins mee eens (5)	Eens (6)	Volledig mee eens (7)
Deze chatbot toont een hoog niveau van kennis. (1)	○	○	○	○	○	○	○

De chatbot zal in staat zijn om complexe vragen te verwerken. (2)	○	○	○	○	○	○	○
De chatbot heeft een hoog niveau van expertise. (3)	○	○	○	○	○	○	○
De chatbot heeft alle up-to-date informatie van het bedrijf. (4)	○	○	○	○	○	○	○

In hoeverre ben je het eens met de volgende stellingen over **deze specifieke chatbot**?

	Volledig mee	Oneens (2)	Enigzins mee	Niet eens, niet	Enigzins mee eens (5)	Eens (6)	Volledig mee eens (7)

	oneens (1)		oneens (3)	oneens (4)			
Ik voel me gerust dat wanneer ik op deze chatbot vertrouw, ik de juiste antwoorden zal krijgen. (1)	○	○	○	○	○	○	○
Deze chatbot zal zeer betrouwbaar zijn. Ik kan erop vertrouwen dat het altijd correct is. (2)	○	○	○	○	○	○	○
Ik ben terughoudend tegenover deze chatbot. (3)	○	○	○	○	○	○	○

Ik vertrouw het advies van deze chatbot. (4)	○	○	○	○	○	○	○

End of Block: Expertise and trust

Start of Block: Instruction 1

Lees de volgende instructie goed door: Je gaat op de volgende pagina gebruik maken van deze chatbot om informatie op te zoeken voor een vlucht naar **Los Angeles**. Je wilt graag op **1 december 2024** vertrekken. Ga in gesprek met de chatbot om de volgende informatie te vinden: **- Wat zijn de kosten voor een vliegticket naar Los Angeles? - Wat zijn de kosten om een extra koffer te voegen?** Druk op enter op je toetsenbord om je chatbericht te versturen. De chatbot zal je door het gesprek leiden en aangeven wanneer je verder kan gaan met de enquête voor de laatste vragen.

End of Block: Instruction 1

Start of Block: Chatbot condition 2- AI- succesful

Wat zijn de kosten voor een vliegticket naar Los Angeles? (1 december) - Wat zijn de kosten om een extra koffer toe te voegen?

[Chatbot frame]

End of Block: Chatbot condition 2- AI- succesful

Start of Block: Instruction 2

Lees de volgende instructie goed door: Je gaat op de volgende pagina gebruik maken van deze chatbot om informatie op te zoeken voor een vlucht naar **Los Angeles**. Je wilt graag op **1 december 2024** vertrekken. Ga in gesprek met de chatbot om de volgende informatie te vinden: **- Wat zijn de kosten voor een vliegticket naar Los Angeles? - Wat zijn de kosten om een extra koffer toe te voegen?** Druk op enter op je toetsenbord om je chatbericht te versturen. De chatbot zal je door het gesprek leiden en aangeven wanneer je verder kan gaan met de enquête voor de laatste vragen.

End of Block: Instruction 2

Start of Block: Chatbot condition 1 - AI- error

Wat zijn de kosten voor een vliegticket naar Los Angeles? (1 december) - Wat zijn de kosten om een extra koffer toe te voegen?

[Chatbot frame]

End of Block: Chatbot condition 1 - AI- error

Start of Block: Instruction 3

Lees de volgende instructie goed door: Je gaat op de volgende pagina gebruik maken van deze chatbot om informatie op te zoeken voor een vlucht naar **Los Angeles**. Je wilt graag op **1 december 2024** vertrekken. Ga in gesprek met de chatbot om de volgende informatie te vinden: **- Wat zijn de kosten voor een vliegticket naar Los Angeles? - Wat zijn de kosten om een extra koffer toe te voegen?** Druk op enter op je toetsenbord om je chatbericht te versturen. De chatbot zal je door het gesprek leiden en aangeven wanneer je verder kan gaan met de enquête voor de laatste vragen.

End of Block: Instruction 3

Start of Block: Chatbot condition 3- NonAI-failure

Wat zijn de kosten voor een vliegticket naar Los Angeles? (1 december) - Wat zijn de kosten om een extra koffer toe te voegen?

[Chatbot frame]

End of Block: Chatbot condition 3- NonAI-failure

Start of Block: Instruction 4

Lees de volgende instructie goed door: Je gaat op de volgende pagina gebruik maken van deze chatbot om informatie op te zoeken voor een vlucht naar **Los Angeles**. Je wilt graag op **1 december 2024** vertrekken. Ga in gesprek met de chatbot om de volgende informatie te vinden: **- Wat zijn de kosten voor een vliegticket naar Los Angeles? - Wat zijn de kosten om een extra koffer toe te voegen?** Druk op enter op je toetsenbord om je chatbericht te versturen. De chatbot zal je door het gesprek leiden en aangeven wanneer je verder kan gaan met de enquête voor de laatste vragen.

End of Block: Instruction 4

Start of Block: Chatbot condition 4 - NonAi - succesful

Wat zijn de kosten voor een vliegticket naar Los Angeles? (1 december) - Wat zijn de kosten om een extra koffer toe te voegen?

[Chatbot frame]

End of Block: Chatbot condition 4 - NonAi - succesful

Start of Block: Trust and expectancy violations

Na je gesprek met deze chatbot, in hoeverre ben je het eens met de volgende stellingen?

	Volledig mee oneens (1)	Oneens (2)	Enigzins mee oneens (3)	Niet eens, niet oneens (4)	Enigzins mee eens (5)	Eens (6)	Volledig mee eens (7)
Ik voel me gerust dat wanneer ik op deze chatbot vertrouw, ik de juiste antwoorden krijg. (1)	○	○	○	○	○	○	○
Deze chatbot was zeer betrouwbaar. Ik kan erop rekenen dat het altijd correct is. (2)	○	○	○	○	○	○	○
Na het gebruik, ben ik terughoudend tegenover	○	○	○	○	○	○	○

deze chatbot. (3)

Ik vertrouw het advies van deze chatbot. (4)

Vergelijk je ervaring met de chatbot, met wat je had verwacht na het zien van het **introductiebericht:**

	Volledig mee oneens (1)	Oneens (2)	Enigzins mee oneens (3)	Niet eens, niet oneens (4)	Enigzins mee eens (5)	Eens (6)	Volledig mee eens (7)
De chatbot voldeed aan mijn verwachtingen. (1)							
De chatbot gedroeg zich anders dan verwacht. (2)							
Ik had de capaciteiten van de chatbot							

hoger ingeschat. (3)							
De interactie met de chatbot heeft mij teleurgesteld. (4)	○	○	○	○	○	○	○

End of Block: Trust and expectancy violations

Start of Block: Last questions

Is het gelukt om in je gesprek op **beide** vragen het juiste antwoord te krijgen van de chatbot?

○ Ja (1)

○ Nee (2)

Heb je wel eens gebruik gemaakt van een dergelijke **customer service chatbot** op een website?

○ Ja (1)

○ Nee (2)

○ Weet ik niet zeker (3)

Display This Question:

If Heb je wel eens gebruik gemaakt van een dergelijke customer service chatbot op een website? , Ja Is Displayed

Hoe vaak maak je doorgaans gebruik van een customer service chatbot op een website?

- ○ Minder dan eens per jaar (5)
- ○ Enkele keren per jaar (1)
- ○ Maandelijks (2)
- ○ Wekelijks (3)
- ○ Meerdere keren per week (4)

Ben je bekend met de term **ChatGPT**?

- ○ Ja (1)
- ○ Nee (2)
- ○ Weet ik niet zeker (3)

Hoe vaak maak je doorgaans gebruik van een **Artificial Intelligence (AI)** tool zoals ChatGPT?

- ○ Minder dan eens per jaar (1)
- ○ Enkele keren per jaar (2)
- ○ Maandelijks (3)
- ○ Wekelijks (4)

○ Meerdere keren per week (5)

Hier volgen de laatste vragen over jouw **algemene** mening over **AI** tools zoals ChatGPT:

	Volledig mee oneens (1)	Oneens (2)	Enigzins mee oneens (3)	Niet eens, niet oneens (4)	Enigzins mee eens (5)	Eens (6)	Volledig mee eens (7)
Ik ben onder de indruk van wat AI kan doen. (2)	○	○	○	○	○	○	○
Ik denk dat AI-systemen veel fouten maken. (3)	○	○	○	○	○	○	○
Er zijn veel nuttige toepassingen van AI. (7)	○	○	○	○	○	○	○
Een AI-system zou beter zijn dan een	○	○	○	○	○	○	○

werknemer

in veel

routine

taken. (8)

End of Block: Last question

Start of Block: End of survey

Dit is het einde van de survey. Je antwoorden zijn opgeslagen. Bedankt voor je deelname!

Dit onderzoek ging over de invloed van het benoemen van Generatieve Kunstmatige Intelligentie (ChatGPT) op het vertrouwen en de indruk van expertise van chatbot gebruikers. De interactie met de chatbot leidde bij sommige participanten bewust tot een fout in het gesprek om teleurstelling te meten. Het doel van dit onderzoek is om ervaringen met chatbots te verbeteren en transparantie te verhogen.

Vond je dit onderzoek leuk? Stuur het dan gerust door naar vrienden en familie! Ik heb 160 deelnemers nodig voor duidelijke inzichten dus alle hulp wordt enorm gewaardeerd.

Voor SurveyCircle-gebruikers (www.surveycircle.com): De Survey Code is: 5YAX-XLH6-3QYD-Z573
Wissel Survey Code in met één klik: https://www.surveycircle.com/5YAX-XLH6-3QYD-Z573/

Voor Serveyswap-gebruikers (SurveySwap.io) ga naar: https://surveyswap.io/sr/FVZC-JD46-

WNXS

Of vul de code handmatig in: FVZC-JD46-WNXS

Link to the survey preview

https://tilburghumanities.eu.qualtrics.com/jfe/preview/previewId/9ba2faaf-cff4-4cd7-89bc-129df6cd7f85/SV_cCFUwn2Rl8xawPs?Q_CHL=preview&Q_SurveyVersionID=current