# Explaining the Wait

## How Justifying Chatbot Response Delays Impact User Trust

Zhang, Zhengquan; Tsiakas, Konstantinos; Schneegass, Christina

**DOI**
10.1145/3640794.3665550

**Publication date**
2024

**Document Version**
Final published version

**Published in**
CUI '24

# *Explaining the Wait:* How Justifying Chatbot Response Delays Impact User Trust


Zhengquan Zhang
mitterchar@gmail.com
TU Delft
Delft, Netherlands

Konstantinos Tsiakas
K.Tsiakas@tudelft.nl
TU Delft
Delft, Netherlands

Christina Schneegass
c.schneegass@tudelft.nl
TU Delft
Delft, Netherlands



## ABSTRACT

In human communication, responding to a question very slowly or quickly influences our trust in the answer. As chatbots evolve to increasingly mimic human speech, response speed can be artificially varied to create certain impressions on users. However, studies remain inconclusive, potentially due to the absence of contextual cues that allow for interpretation of the delay. Thus, this study explores textual explanations that justify the *instant* and *dynamic* – dependent on answer length – response delays. We derive five design variations based on prior work and evaluate their impact on the chatbot's perceived social presence and transparency ($N = 10$). In a between-subject online study ($N = 194$), we then evaluate the influence of the highest-rated justification on users' perceptions of chatbot transparency, social presence, and trust for the two delay conditions. Results demonstrate that while such justifications enhance perceived transparency and trust in the immediate response scenario, they show no effect in the dynamic delay context.



## CCS CONCEPTS

• **Human-centered computing** → Natural language interfaces; **Empirical studies in HCI**.

## KEYWORDS

Explainability, Chatbot Response Delay, Transparency, Social Presence, Trust

**ACM Reference Format:**
Zhengquan Zhang, Konstantinos Tsiakas, and Christina Schneegass. 2024. *Explaining the Wait:* How Justifying Chatbot Response Delays Impact User Trust. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3640794.3665550


## 1 INTRODUCTION

With the rapid advancements in chatbot performance led by OpenAI's ChatGPT[1], conversational user interfaces (CUIs) are now able to provide answers to almost any question, showing their skills in many application contexts, from customer service (e.g., [19, 69]) to healthcare support (e.g., [18, 34]). Yet, due to the complexity of artificial intelligence (AI) and its underlying neural network(s), the process of how the chatbot interprets the input questions and generates the answer is often unpredictable and unexplainable. Having no justification or source for the presented output, especially given the frequent occurrence of hallucinations [2, 33], can impact users' trust in the system.

Given that chatbot interactions are becoming increasingly natural and people tend to anthropomorphize communication with their digital conversation partners [39, 63], research is looking more closely at theories of how trust is established in human-to-human communication. Among many variables of conversations that impact our trust in someone's answer is a person's response speed. For example, quick responses might be interpreted as confidence or eagerness, while delayed responses might hint at a careful answer generation but also potentially uncertainty. However, unlike humans, chatbots don't 'think' or 'contemplate'. A delay in their response doesn't necessarily signify an internal cognitive process but instead could be caused by computing time or network speed. In this work, we look at response speed as a dedicated User Interface (UI) design choice. Research has found that instant responses reduce the feeling of a natural conversation and decrease user satisfaction [28]. Other studies found that dynamically adapted responses that consider the length of the answer can increase social presence and trustworthiness [9, 28]. In contrast to human-to-human communication, during which we use additional cues such as facial expressions or body posture to interpret a response delay, human-chatbot interaction lacks contextual cues, making it challenging for users to interpret the speed of the response, potentially further decreasing their trust in them.

To reduce the uncertainty that arises from the lack of contextual cues, this work proposes to employ textual explanations for response delay to create user trust. We will call these explanations "justifications" throughout the paper to deliberately distance our work from how the term "explanations" is used in the Explainable AI (XAI) community. We consider our justifications a special case of explanation, in which the provided statements do not necessarily reflect the exact mechanisms behind an AI black box but rather aim to justify a certain observable behavior.

In XAI, explanations have been shown to increase trust and transparency in certain black-box systems, such as the algorithm controlling the Facebook news feed [54] or recommendations of a digital fitness coach [15]. Yet, as of today, research in CUIs has not yet looked into using explanations or justifications as a means for addressing response delay. For other facets of CUIs, explanations revealed mixed effects on user trust, thus emphasizing the need for further in-depth research. For example, studies on human-agent

[1] ChatGPT – https://chat.openai.com/, last accessed February 26th, 2024




*CUI '24, July 08–10, 2024, Luxembourg, Luxembourg*



ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665550

interaction (cf. [4, 47]) have shown that transparency and trust depend on task, context, and the actual explanations provided.

For this study, we mimic the response behaviour of (a) retrieval-based and (b) generation-based chatbots. The first type already has a fitting answer stored in its knowledge base because it was trained for a specific purpose and is thus able to provide an almost instant response. In contrast, generation-based chatbots craft a response specifically for the user's question and thus show a dynamic response delay dependent on the answer length. The goal of this work is to investigate the effect of response delay and response delay justifications in these two conditions on users' perception of social presence, subjective transparency, and, ultimately, trust in the system. For this purpose, we review relevant prior literature and generate a set of five justifications for each chatbot category. We evaluate these justifications in a preliminary study ($N = 10$) regarding the extent to which they create a feeling of social presence and subjective transparency, as well as overall suitability and user preferences. The highest-rated justification is then used in an online between-subject user study ($N = 197$), in which participants rate their experience with interacting with the two response delay conditions (instant response for the retrieval-based chatbot, dynamic delay for the generative response chatbot), with and without justifications ($2 \times 2$ design). The results show that justifications of response delay can positively influence the perceived transparency of the chatbot and improve the trust in the retrieval-based chatbot with the instant response while minimally lowering trust in the generation-based condition that deployed a dynamic delay.

Overall, this paper contributes five textual justifications for instant and dynamic response delays designed to foster social presence and subjective transparency. Furthermore, we provide empirical evidence of the positive effect of justifications on transparency as well as on trust in the instant response delay condition. We conclude by summarizing design implications for utilizing response delays and response delay justifications toward trust-enhancing chatbots.

## 2 RELATED WORK

### 2.1 Chatbot Taxonomies and Classifications - Retrieval-based vs. Generative Chatbots

Chatbot-based systems can be categorized by application context, technical aspects, and design elements. A technical review for modern chatbot systems categorizes them further based on chatbot knowledge and means of response generation [45]. Chatbot knowledge is divided into open and closed domain systems. In closed-domain systems, chatbots are designed to support a task-oriented interaction and require a task-specific knowledge base, thus following rule-based and template-based methods. On the other hand, open-domain chatbots require a general knowledge base and the ability to handle free-form conversations [7]. They apply data-driven approaches to tackle issues of natural language understanding and generation that can occur due to the diversity of possible dialogue topics and natural language [60].

These two approaches are the two primary means of answer generation, which we will focus on in this paper: **retrieval-based** and **generative** [68]. Retrieval methods refer to the process of selecting the best output from shortlisted candidates based on the user's input, while generative methods deploy techniques for output generation based on trained classifiers. Other classifications also include rule-based as an additional category of chatbots, which can describe chatbot systems that do not deploy a machine learning (ML) algorithm to retrieve or generate a response [1]. Retrieval-based chatbots are considered reliable within their knowledge domain since the responses are based on a predefined knowledge base and are less likely to generate inappropriate or irrelevant responses. However, recent natural language processing (NLP) advancements have led to a shift toward generation-based models [46]. Generation-based chatbots have the advantage of being more flexible and able to generate responses that are personalized instead of solely predefined. Due to the two means of response selection and presentation, the speed of these actions might vary. For example, long and complex answers could be presented faster in a predefined rule-based system as compared to a generative system.

### 2.2 Trust in Chatbots

An emerging theme of human-chatbot interaction identified in the analysis of Rapp et al. [56] is the acceptance of the chatbot during the interaction, related to aspects of expectation, engagement, and trust. Trust is essential in human-chatbot interaction and relates to the user's perception and expectation of the chatbot functionality and performance [73]. Research studies aim to identify ways to quantify and predict trust in chatbot-based systems and recognize factors such as credibility, anthropomorphism, and social presence [26, 72]. Given that building trust in complex technical systems does not only depend on the design of the technical system [42], there is the need for more research into systematically deriving knowledge and design principles for trust-enhancing features, including transparency and social presence [76], which we will discuss in more detail in the following. In regard to this paper's topic of response delay, the concepts of humanness and social presence are especially vital, as human response generation is inherently different from technologies. The following sections will discuss the commonalities and differences in more detail.

### 2.3 Social Presence

Human interactions are rich with social cues, from facial expressions to hand gestures [6]. Building on the notion that technology can mimic these cues, the Social Response Theory (SRT) posits that users react to technology similarly than to human-initiated social cues, such as natural language or anthropomorphic appearances [50, 57]. Following the evidence that a human-computer relationship is fundamentally social, research studies focus on how to design social responses through artificial social cues [30, 51]. Human users subconsciously treat socially interactive technology as a social entity, attributing human traits to it, which enhances the perceived social presence [49] – a term coined in 1976 to describe the subjective perception that another person or entity is present [62]. In the domain of e-commerce, social presence is an important aspect since it affects consumers' trust [24]. In implementing SRT, studies have found that human-robot interactions perceived as human-like and spontaneous can elevate social presence and trust [70]. Chatbot research, on the other hand, has mainly focused on verbal (e.g., greetings and humor) and visual cues (e.g.,

avatars) to promote a sense of social presence [3, 14, 30, 48]. The study by Moussawi et al. [48], for instance, found that a chatbot that can respond cheekily and in a humorous way to a question about its name can appear to be more human-like. Additionally, research underscored a chatbot's social presence as a determinant of user trust [75], demonstrating how chatbot anthropomorphism in retail contexts, like using first-person phrasing, enhances user trust and overall satisfaction [39]. Yet, insights on other cues, particularly response time, remain scarce [17].

## 2.4 Response Delay

Response delay is considered a vital social cue in technology-mediated interaction [32, 63]. In chatbots, unlike human counterparts who need time to read a message and enter a response, retrieval-based chatbots can instantly process user input and provide a response [46]. Yet, these instant responses may reduce the feeling of a natural conversation and decrease user satisfaction compared to chatbots that implement a dynamic response delay [28]. In contrast, some studies indicate the opposite, i.e., that chatbots with delayed responses are perceived as less likable [59]. Previous studies on online customer service have found that compared to nearly instant response, the dynamic response delay of the chatbot can improve the perceived social presence and trustworthiness [9, 28]. The diverging findings could be rooted in the fact that chatbot communication delays are subject to users' interpretations. They may be interpreted as a feature or a dysfunction of the chatbot. Thus, our first question (RQ) aims to tackle the diverging results and confirm the existing findings to create a basis for our follow-up research questions:

> **RQ1: How does response delay influence users' trust, subjective transparency, and social presence in chatbots?**

More specifically, this work aims to explore how different types of response delay, namely (a) *near-instant* and (b) *dynamic* response delay, affect users' trust in a chatbot. A near-instant response is a static response for all prompt-response pairs and aims to simulate the interaction with a knowledge-based chatbot, i.e., retrieval-based, while the dynamic response delay takes into consideration prompt and response characteristics, e.g., length of prompt/response, which can simulate the response generation process of a generation-based chatbot. Given the divergent findings of prior literature, we explore the relationship between response delay on trust without expecting a direction of the effect. Furthermore, the rapid progress in chatbot technologies and their increasing ubiquity in people's everyday lives, which influences people's overall level of trust in such systems, demands up-to-date research to confirm the effects.

## 2.5 Transparency and Explanations in Chatbots

With the rise of the big generative language models BERT [13, 35], T5 [55], GPT-3 [5], and others, chatbot systems often appear as a "black box" to the user. They make it difficult to understand the underlying process of input processing and response generation as well as understanding why something did not work and what actions are actually possible [36]. Explanation capabilities can significantly impact user trust while interacting with an intelligent system [27]. Furthermore, explanation-based interactions can also help users learn how they can efficiently interact with chatbots to complete a task, even in the case of a dysfunction [65]. Research in XAI aims to address the black-box problem either through the development of explainable models that can justify their decisions or through visualizations of the underlying AI processes [71].

Since AI has been extensively used for the development of chatbot-based systems, there is a need to integrate explainability methods into the design process as a way to enable users to understand, trust, and manage their interaction with chatbots [21, 41]. A common approach to increase system transparency and users' perception of the system capabilities is through explanation interfaces, designed to communicate the explanations to the user efficiently [15, 54]. Two main questions for designing explanations are: (a) what to explain and (b) how to explain it. Considering the different features of a chatbot-based system, including both algorithmic and design aspects, there are several opportunities for XAI methods to enhance transparency. In the domain of Human-Agent interaction, studies have looked at the effect of transparency on numerous facets of interaction quality, such as operator performance, workload, situational awareness, trust in automation, and perceived usability. Bhaskara et al. [4] summarize the results showing that results are not consistent and hypothesize that this is likely due to the different contextual and task settings as well as dependent on the actual level of transparency provided. For trust, in particular, they report a positive relationship between transparency and trust but mention inconsistencies in the results depending on how transparency is achieved, e.g., by showing system uncertainty or predicted outcomes Bhaskara et al. [4]. Similarly contradicting findings resulted from the study of Lyons et al. [47], who investigated the role of explanations in unexpected actions by a robot partner. They found that while certain explanations do increase trust, others are no better than providing no explanation at all [47].

Furthermore, in the context of a chatbot-based movie recommendation system, the explanation type (i.e., *why* vs. *why-not* explanations) has been shown to have an effect on how users perceive the system's transparency and how much they trust the system's decisions [66]. Another study investigated if transparency and time fillers can create a positive user experience during the waiting time in customer-chatbot conversations [67]. Customers who interacted with a chatbot that provided information about the waiting time and status perceived it as more transparent and reliable. What has not yet been investigated is the use of explanations or justifications for response delay times. Thus, our study aims to address the following research question:

> **RQ2: How does justifying the response delay influence subjective transparency and users' trust in chatbots?**

More specifically, we want to explore whether providing a justification for the response delay and, thus, explaining the chatbot's underlying way of response generation (retrieval-based or generative), affects users' trust. We hypothesize that a chatbot that justifies why its response is instant or delayed will be perceived as more transparent and, thus, as more trustworthy. We expect the justifications to have a direct effect on perceived transparency and an indirect effect on trust. Figure 1 summarizes how we expect the concepts of response delay, justifications, transparency, presence, and trust to be related based on the related work presented above.

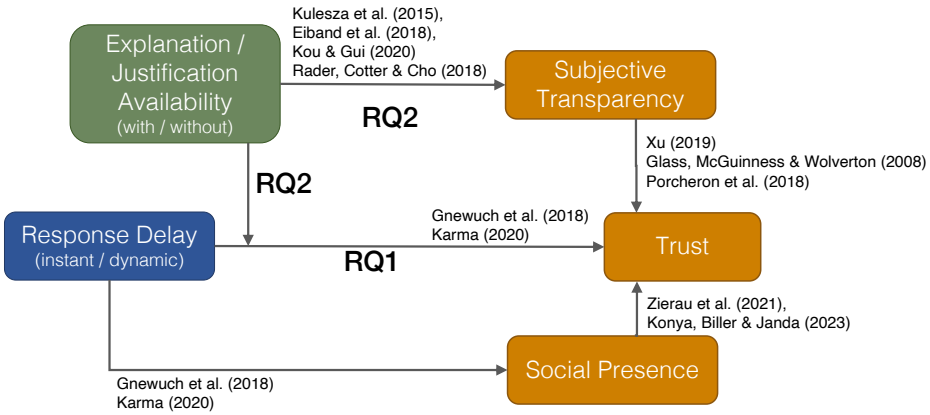



**Figure 1: This research model summarizes the theoretical concepts relevant to this work. Links with references indicate the availability of prior research, while the RQs indicate the open questions our work aims to tackle.**

In this work, we will investigate how justifications affect the perception of response delays in chatbots. For that purpose, we first design different types of textual justifications and evaluate them in an initial user study regarding their effect on the chatbot's social presence and transparency. We will then select the highest-ranking justification and conduct a large-scale online study to investigate its effect on trust and the mediating effect on subjective transparency. We discuss our findings and derive implications for the design and implementation of textual justifications in rule-based and generative chatbot applications.

## 3 JUSTIFICATION GENERATION & EVALUATION

Based on related work, we created two chatbots, one mimicking a retrieval-based approach and one a generation-based approach, as common examples of chatbot types. We understand retrieval-based chatbots as those instantly pulling static responses from a pre-defined knowledge base, while the latter dynamically constructs replies, thus introducing variable response times [46] (see Section 2.1 for more details). In this section, we will report on our process of generating and validating justifications to explain both the instant and dynamic response delay.

### 3.1 Response Delay Design

We created two chatbot prototypes that artificially modulate the response delay to be perceived as either a retrieval-based or generation-based chatbot. Both work on a pre-defined knowledge base to control the content of the conversation in this pretest compared to using an unpredictable Large-Language Model (LLM). The delay time for the retrieval-based chatbot was set to be nearly instant, i.e., we present the justifications after 0.05s and the response after 0.1s to imitate an almost immediate response with only negligible technical delay. The response delay for the generative chatbot was dynamically changed by utilizing a calculation from [31], where the response delay is determined based on the number of characters in the chatbot's response, specifically 50 milliseconds (ms) per character. We also opted to use predefined conversations in the generative bot condition as compared to a real generative bot to ensure that prompts are used as planned by our participants and response time is controlled.

### 3.2 Justification Design

While justifications that indicate query processing and response generation could be designed in any modality, from a visual icon (e.g., three dots to show typing) to an auditory cue, we opted for a textual description. Written text provides the opportunity to unobtrusively communicate detailed information about the chatbot's behaviour and inner workings. At its core, all explanatory text snippets we design are meant to indicate if the chatbot is retrieving an already existing response (retrieval-based chatbot), justifying a very quick response, or generating a response on the spot (generation-based), which will explain response times that become longer with more extensive responses. Table 1 outlines the five distinct response-delay justification styles we designed and tested for each delay condition. The "*Basic*" justification offers only a fundamental description of the chatbot's response mechanism. "*First-person*" deploys a self-referential approach, which refers to the study by Konya-Baumbach et al. [39] who has shown that implementing first-person descriptions in the response can enhance social presence while interacting with a chatbot. The "*Detail*" version goes deeper into the chatbot's operational process, aiming to increase perceived transparency by providing more details about the functionality. Merging these, the "*First-person + Detail*" justification incorporates a detailed description with first-person narration. Lastly, the "*Humor*" version integrates humor to make the chatbot appear more human-like, as prior research suggests its efficacy in augmenting social presence [48][2].

[2] The humerous explanation was generated based on the research teams' collective intuition for humor.

**Table 1: Overview of the ten justification statements generated and evaluated in this work. Five for the retrieval-based (instant response delay) and five for the generation-based (dynamic response delay) condition. The revised justification incorporates findings from the pretest and presents the final text we will continue to use for the main study.**

| Type | Justification retrieval-based (instant) | Justification generation-based (dynamic) |
|---|---|---|
| **Basic** | Retrieving the answer. | Generating the answer. |
| **First-person** | I am retrieving the answer for you. | I am generating the answer for you. |
| **Detail** | Searching in the knowledge base for the answer that matches the question intent most. | Using the question as a prompt to generate the answer from the pre-trained model. |
| **First-person + Detail** | I am searching in my knowledge base for the answer that matches your question intent most. | I am using your question as a prompt to generate the answer from the pre-trained model. |
| **Humor** | I'm like a witty librarian in my knowledge base, searching for the answer that fits your question like a puzzle piece. | I channel my inner word wizardry, conjuring up sentences word by word like a mischievous magician performing a linguistic sleight of hand. |
| **Revised justificationsafter pretest** | | |
| **First-person + Detail** | I am searching for the answer that matches your question intent most in my knowledge base. | I am using your question as a prompt to generate the answer from my latest AI model. |

## 3.3 Prototype

We created two chatbot prototypes[3] to explore the nuances of retrieval-based and generation-based chatbots using Gradio[4], an open-source python library for rapid UI development for ML models. Though Gradio has limitations in detailed UI customization, it met our study's needs to vary the response delay (using Gradio's coding features) and explanatory styles. The retrieval-based prototype was configured for near-instantaneous replies, while the generation-based prototype used a dynamic delay based on the formula of Holtgraves and Han [31] – 50 ms per response character. While Gradio can be integrated with real-time LLMs like ChatGPT, we preloaded answers for our controlled test setting, avoiding the unpredictability of real-time LLMs. We also standardized questions and responses to minimize any variables affecting user-chatbot interactions. By pressing a button, participants could preselect a question from a set, focusing on topics such as sustainability and history. We categorized questions into four groups based on their length and the length of their corresponding answers. Short questions contained less than ten words, and long questions around 20. Short responses contained less than 25 words, whereas long responses contained around 50 words. Both chatbot prototypes, instant and dynamic response, display three dots commonly used as typing indicators when generating the answer (displayed after the textual justification). The chatbot prototypes are showcased in Figure 2, exemplarily showing the retrieval-based chatbot with the basic justification (Figure 2a) and the generation-based chatbot with the detailed justification (Figure 2b).

## 3.4 Procedure

Before initiating this pretest, we secured ethical clearance from TU Delft's Human Research Ethics Committee (HREC, case ID 3181). During the pretest phase, all participants were informed about the study process, data management, and anonymization policy and gave informed consent. The pretest was conducted as a within-subject experiment. Participants engaged with chatbots at first in a no-justification control condition. Afterwards, we led them through all five justification scenarios described in Table 1 (i.e., (1) *Basic*, (2) *First-person*, (3) *Detail*, (4) *First-person* combined with *Detail*, (5) *Humor*) and two delay types (instant and dynamic), resulting in 10 distinct conditions. We provided the definitions of social presence[5] and subjective transparency[6] to participants and provided necessary explanations if the concepts were unclear. We decided against counterbalancing and presented the justification in order of increasing complexity as listed above to avoid bias and tasked participants with ranking them concerning perceived social presence and transparency, from lowest rank (5) to highest rank (1). Conclusively, we conducted a brief interview to capture their feedback on response delay justification clarity, comprehensibility, and other related factors. We analyzed the interviews using a light version of a thematic analysis to identify themes as well as shared and divided opinions on the justifications in the interview results.

## 3.5 Sample and Recruitment

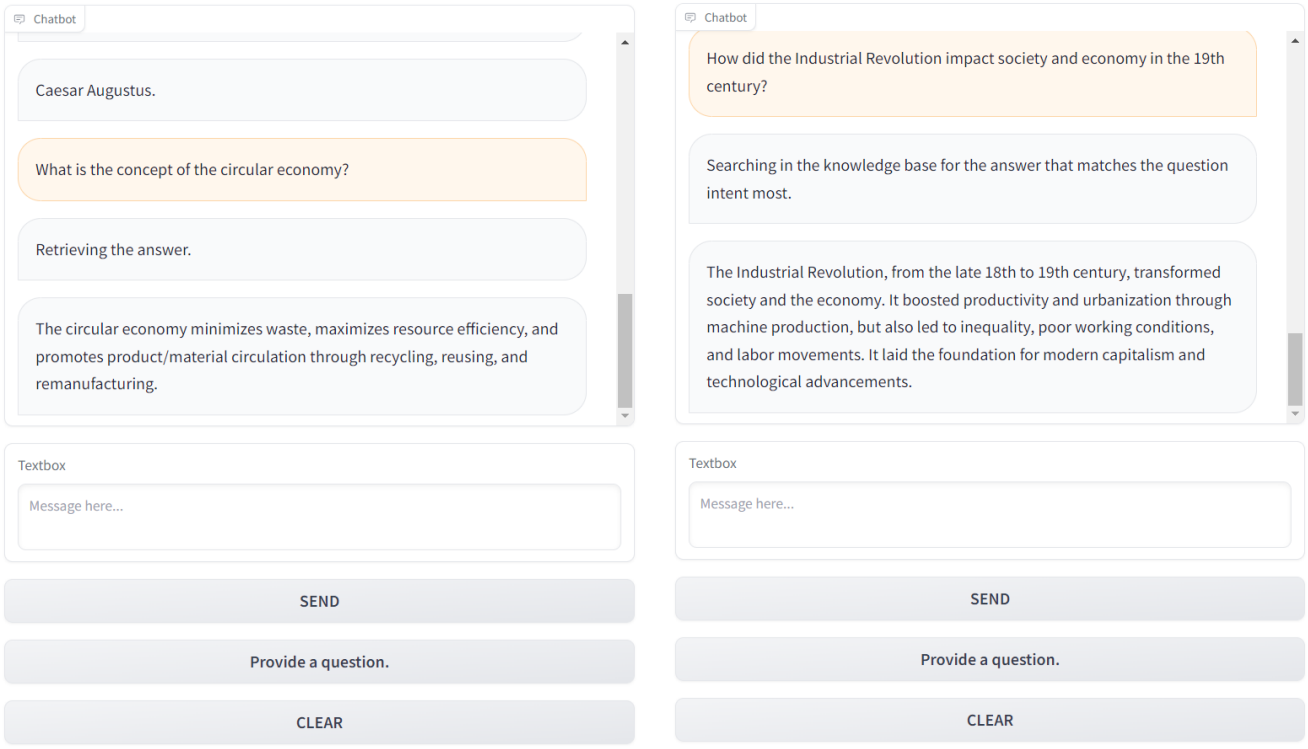
We recruited ten participants (five identifying as male, five identifying as female) on a voluntary basis (no compensation) from our university. All participants were master's level students in engineering, aged 24 to 26 ($M = 25$, $SD = .82$). They were second-language English speakers but with advanced education (minimum bachelor's degree), through which we expect sufficient English language proficiency for this study. Every participant reported previous experience with chatbots, predominantly using ChatGPT regularly in recent months.

[3] GitHub repository pre-test chatbot prototypes: https://github.com/MitterChar/Chatbot-prototype

[4] Gradio – https://www.gradio.app/, last accessed February 26th, 2024

[5] Definition of Social Presence by Short et al. [62]: "*the degree of salience of the other person in a mediated communication and the consequent salience of their interpersonal interactions*". The definition was adjusted over time, highlighting the ability of communication media to express sensations of warmth, personal connection, and sociability, without the need for direct human-to-human contact (cf. [24]).

[6] Definition of Subjective Transparency: In the context of the advice-giving system, Zhao et al. [74], distinguishes between objective transparency and subjective transparency and define "***objective transparency*** *as the extent to which systems release information about how they work, and* ***subjective transparency*** *as the extent to which users perceive such information is available*".




**(a) *Basic* justification** **(b) *Detail* justification**

**Figure 2: Screenshots of the retrieval-based chatbot prototype for two response delay justification conditions, *Basic* (left) and *Detail* (right). User input is shown in orange and chatbot responses are in grey.**

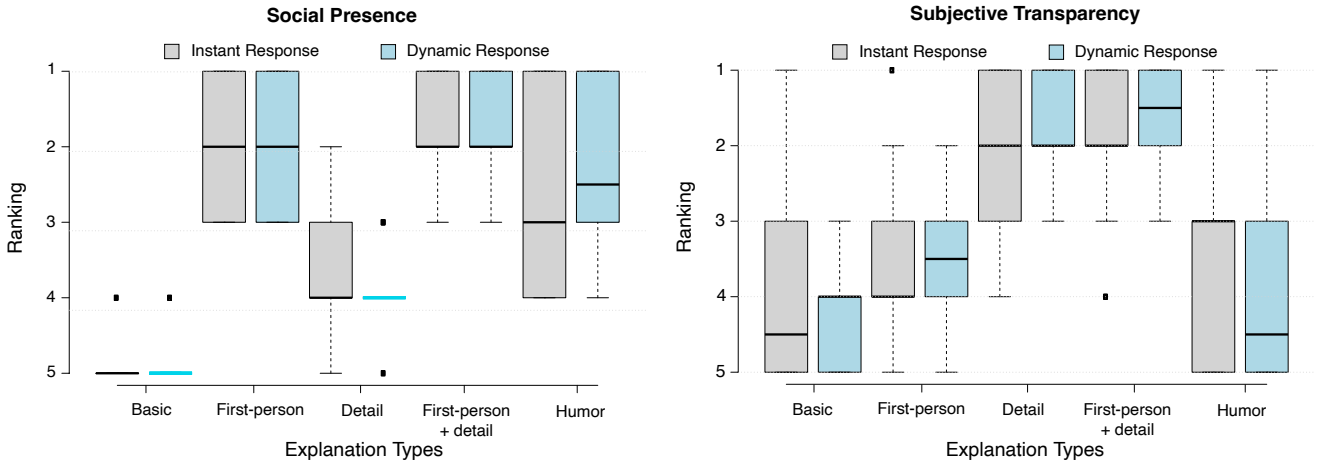
## 3.6 Results and Discussion

The average user rankings revealed that justifications using a first-person perspective (*First-person*, *First-person + Detail*) received the highest ratings on the social presence scale with a median ranking of $Md = 2$ for both conditions for instant and dynamic delay (see Figure 3b). For subjective transparency, the detailed conditions (*Detail* and *First-person + Detail*) were ranked highest with a median of $Md = 2$ for *Detail* (instant and dynamic), and a Median of $Md = 2$ for *First-person + Detail* in the instant and $Md = 1.5$ in the dynamic delay condition (see Figure 3a).

As reasoning for the high social presence rating of the *First-person* conditions, P3, P4, P7, P8, and P10 mentioned that the descriptions "I'm" and "for you" were warmer and could reduce social distance, providing a stronger feeling of social presence. These descriptions made them believe there was an empathic technology helping them. At the same time, the plain justifications (*Basic*) felt like a machine (P1, P3-6). Regarding subjective transparency, participants thought detailed justifications contained more information, thus improving subjective transparency on the workings of the chatbot. When comparing the justification type *Detail* and *First-person + detail*, participants mentioned that again, the description "I'm" made them feel as if the chatbot was more transparent because it provided a vivid scenario in which there is a robot searching for or generating the answer for them (P4, P5). For example, P5 mentioned, "[the chatbot] feels like a person [who] sits there for you, you give him/her orders, and then he/she finds the answer." In addition, some participants thought the *Basic* justification made the chatbot more transparent, but others did not appreciate the objective style of talking. For example, P3 mentioned, "[...] I don't think a cold machine is more transparent". For the humorous justification, participants felt the chatbot had a strong personality but was joking too much, making them feel like it was fooling them, trying to hide the truth (P3) or was not sincere (P4, P8, P9). For example, P3 mentioned, "[...] it's so humorous, I always feel like it's hiding something", and P8 states that "[...] (the *Detail* justification) is more sincere and true compared with [the *Humor* justification]". Thus, the humorous justification did not perform well regarding subjective transparency. The humorous way of talking made participants feel as if the chatbot was trying to make itself perform like a human but not close enough to pass as a human. This also reduced the feeling of social presence (P9). Since the humor justification was more creative than the other justifications, we consider this a limitation of our setup.

## 3.7 Optimizations and Main Takeaways

The pretest aimed to derive a set of textual justifications for response delays caused by retrieval-based and generation-based chatbots that can create a feeling of social presence and transparency. In




**(a) Social presence**

**(b) Subjective transparency**

**Figure 3: User ranking of our justification types' social presence and subjective transparency, distinguished by the two conditions instant and dynamic delay.**

the pretest, the *First-person + Detail* justification performed best regarding social presence and subjective transparency aspects. Thus, we selected the *First-person + Detail* justification for further evaluation in our main test. Secondly, we optimized the statement based on the feedback from the pretest. On the one hand, participants cared about where the chatbot's answer came from – especially when interacting with the generation-based (dynamic delay) chatbot. They cared about the timeliness/version of the underlying model (P5), i.e., ensuring the chatbot is up-to-date. Furthermore, the words "pre-trained model" in the *Detail* justification confused at least one user (P2) and reduced the feeling of social presence in another (P10). Thus, we replaced the phrase "generate the answer from the pre-trained model" in the *First-person detail* justification of the generation-based chatbot with the phrase "generate the answer from my latest AI model". To match the sentence structure of the generation-based chatbot, we revised the word order of the retrieval-based chatbot's justification as well (see Table 1). Thirdly, eight out of ten participants felt the response delay for the dynamic delay condition was too long. Hence, we adjusted our way of calculating response delays using the adapted equation proposed by Gnewuch et al. [28], which will be explained in more detail in the main test section. Lastly, regarding the interface design, participants felt the distinction between the justification text and the actual answer was not obvious and created confusion (P2, P4, P8, P9). Thus, we will also adjust the interface, specifically the way of presenting the justifications.

# 4 MAIN STUDY

## 4.1 Experimental Setup

To investigate our research questions on the influence of response delay and response delay justifications on users' trust in chatbots, we conducted a second study, which took place in July 2023. In this study, participants experienced chatbots in two response delays (instant vs. dynamic), with and without response delay justifications, i.e., following a $2 \times 2$ design, resulting in four study conditions. In a between-subject design, participants randomly interacted with one of the four conditions and evaluated their experience concerning perceived social presence, subjective transparency, and trust toward the chatbot (dependent variables). As described in the related work section, research has already proven certain relationships among response delay, justifications, subjective transparency, social presence, and trust (cf. our research model in Figure 1. In this study, we aim to confirm these relationships and further expect to find that our manipulated response delays and newly introduced justifications create the following effects:

**H1a Justifying the chatbot's response delay will increase the subjective transparency of the chatbot.**

With this hypothesis, we aim to test if the justifications we designed actually make the chatbot's workings more understandable (cf. RQ2). Specifically, we expect the justifications to indicate to the user if the chatbot follows a retrieval-based or generative approach. By doing so, users receive an explanation of why instant responses are provided so quickly and why dynamic responses take longer, depending on the answer length. Thus, the chatbot will, in both conditions, become more transparent and, with this, also more trustworthy with the justifications.

**H1b Justifying the chatbot's response delay will moderate the effect between response delay and trust.**

Based on prior work by Gnewuch et al. [28] and Choedak [9], we expect that users will trust the chatbot more when using a dynamic response delay than an instant response delay (cf. RQ1). We expect justifications to even out this difference, as they help to justify the quick response time. Thus, we expect justifications to increase users' trust more in the instant than in the dynamic condition (cf. RQ2).

## 4.2 Prototype

The chatbot used in this main test is an optimized version of the pretest chatbot (see Figure 4). We optimized the response delay and justifications as explained below (see also Subsection 3.7 for an in-depth justification of our changes).

*4.2.1 Response Delay.* To optimize the response delay for the dynamic delay condition, we now consider the calculation by Gnewuch et al. [28], which is based on the sentence complexity and response length. The complexity (C) of the language used in each message is aligned with the formula of Kincaid et al. [37] for one sentence:

$$C(s) = 0.39\times(\text{total words})+11.8\times(\text{total syllables/total words})-15.59 \quad (1)$$

The complexity values can range from -3.40 to positive infinity. Based on this, the time delay (D) was calculated in seconds based on the complexity value (C(s)) of a sentence:

$$D(s) = 0.2 \times ln(C(s) + 0.5) + 0.5 \text{ for } C(s) > 0 \quad (2)$$

In addition to this delay, we must add a minimum data transmission time of 200 to 400 ms, similar to current real-world chatbot applications [29]. We must consider an estimate of 100 ms for the technical delay caused by Gradio in selecting and printing the response. Thus, the final calculations of the total delays are:

> **Total Delay Static** = Internet delay + Technical delay
> **Total Delay Dynamic** = SUM(D(s)) + Internet delay + Technical delay

For an example chatbot response from the study[7], given a 200ms transmission time, the total delay **static** would be 0.3 seconds, while the total delay **dynamic** would add up to 3.877 seconds.

*4.2.2 Revised Justifications for Response Delay.* As a result of the pretest, we generate a revised set of textual justifications for the instant and dynamic response delay conditions:

> **Instant Delay:** "I am searching for the answer that matches your question intent most in my knowledge base."
> **Dynamic Delay:** "I am using your question as a prompt to generate the answer from my latest AI model."

The justifications themselves were sent only with the unavoidable technical delay by Gradio and appeared immediately in the conversation window.

*4.2.3 Material.* The pretest showed that participants engaged with the chatbot by posing queries unrelated to the subject. To control users' experience in this main study, we utilized pre-recorded videos that depicted user chatbot interactions as our experimental stimuli. Participants were instructed to view these videos and follow the conversation closely. While this procedure improves our control over the experiment, it potentially influences users' experience. We will elaborate on potential effects in the discussion section.

The focus of the chatbot conversation we recorded was on astronomy. In contrast to the themes in the pretest, history and sustainability, we expect less familiarity with the subject. This was meant to hinder participants in assessing the chatbot's response quality and accuracy, thereby focussing their evaluation of its trustworthiness on the interaction and justifications. The main experiment involved five rounds of recorded interaction between the user and chatbot, including starting and concluding messages and three astronomy-related queries, thus resulting in videos between 1:12 min (minimum, instant no justification) and 1:31 min (maximum, dynamic with justification). A sample interaction is illustrated in Figure 4[8]. To generate these texts, we used ChatGPT version 3.5. We used a script to automate the typing and sending of messages to generate the video recording of the dialogues. The time interval between each keystroke was randomized between 20 ms and 100 ms to imitate the irregularities in human typing. The source code for the main test chatbot prototype and the auto-filling script is publicly accessible on Github[9].

## 4.3 Procedure

For our study, we utilized our University's instance of the survey platform Qualtrics[10] integrated with Prolific[11] for participant recruitment. Participants were informed about the study data management (ensuring GDPR) and provided informed consent to proceed[12]. They were informed about the study's general purpose (to evaluate the interaction design of the presented chatbot) and procedure and randomly assigned to one of the four experimental conditions. Subsequently, participants were asked to attentively watch the video of the recorded chatbot interaction of the respective condition and responded to a set of questionnaires (consistent across conditions). To ensure participants' attention, we included control questions (e.g., "I swim across the Atlantic Ocean to commute daily"). The expected answer was a (strong) disagreement, indicating that participants were paying attention. We also inserted two questions related to the video's content to check if participants attentively followed the conversation displayed in the video. All participants successfully answered these questions. The study started with the video stimuli to ensure high attention levels. Subsequently, we provided several questionnaires and assessed demographic information. The questionnaire list included: perceived delay manipulation check [22], social presence [23], subjective transparency [74], trust [8], chatbot experience [29], astronomy proficiency self-assessment, and the Affinity for Technology Interaction (ATI) scale [20] (full list in Table 3 in Appendix). To maintain confidentiality, we fully anonymized all data and assigned random participant IDs.

A trial run suggested that participants would take, on average, six minutes to finish. In line with Prolific's fair payment guidelines, which recommend a minimum of USD ($) 6.50/h, we decided on a $1.25 incentive for the whole study. The median time for completing the survey, as reported by Prolific after study completion, was 06:01 minutes, resulting in an average hourly compensation of $12.50.

[7] Example chatbot response from the main study: "A supernova is a large explosion that takes place at the end of a star's life cycle. It occurs when there's a change in the core, or center, of a star. This change can happen in two different ways, with both resulting in a supernova. It's the largest explosion that takes place in space and can shine as brightly as an entire galaxy of billions of normal stars."

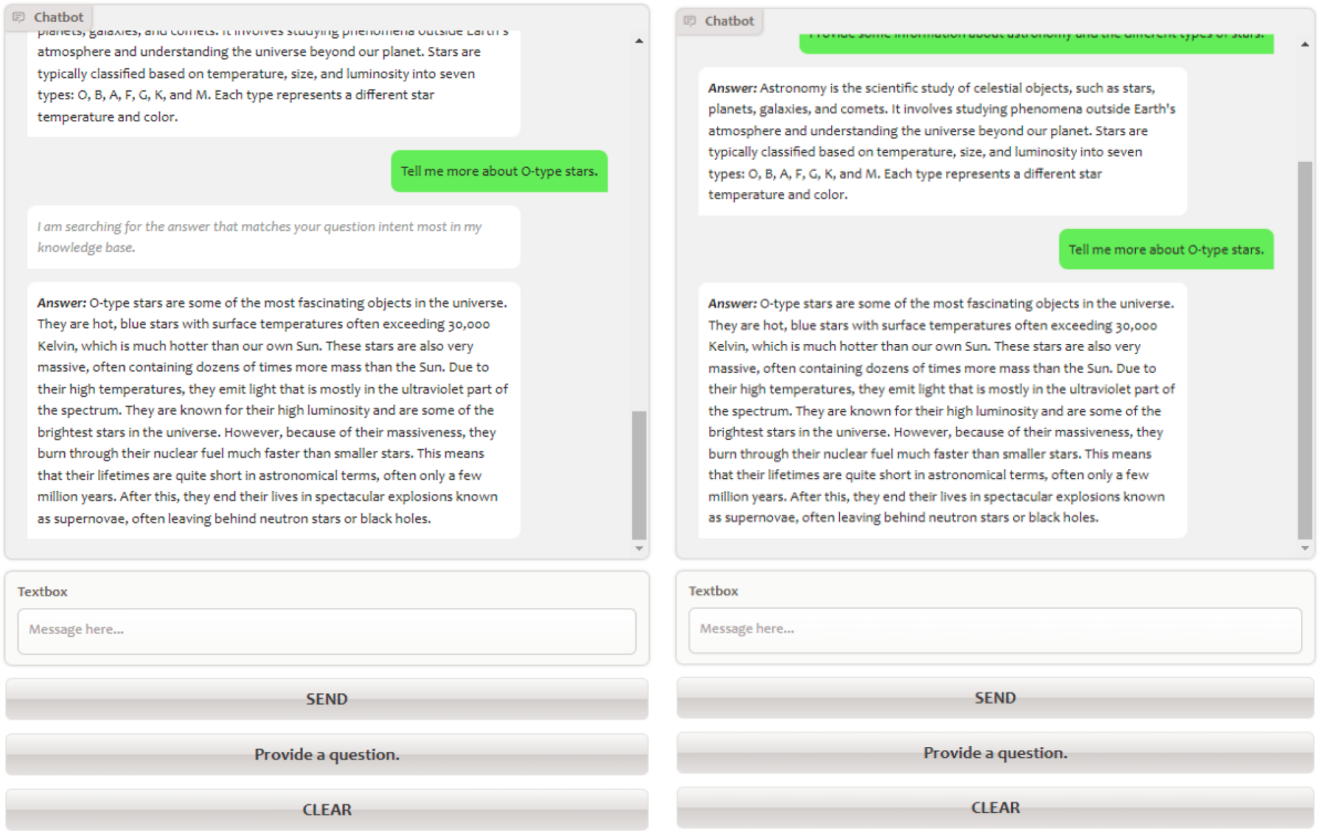
[8] The complete conversation flows for all conditions are provided as complementary material.

[9] GitHub Repository main study chatbot prototype: https://github.com/MitterChar/Chatbot-prototype

[10] Qualtrics – https://www.qualtrics.com/ – last accessed February 26th, 2024

[11] Prolific – https://www.prolific.co/

[12] We received ethical approval by TU Delft's human research ethics committee (HREC) before the start of the study, case ID 3181.




**(a) Experimental Condition (with justification)**

**(b) Control Condition (no justification)**

**Figure 4: Chatbot interfaces for the two conditions in our study.**

## 4.4 Participants & Sample Size Calculation

We recruited participants through *Prolific*. Respondents were prompted to complete the study using a PC or laptop, as the text in the chatbot conversation in the videos might become hard to read on mobile devices. An apriori power analysis conducted with G*Power [16] indicated that at a 0.05 significance level, we needed a minimum of 179 respondents to attain a statistical power of 0.80 in a non-parametric Mann-Whitney U tests, assuming a medium effect size ($f = 0.25$). For an ANOVA with main effects and interactions, attaining medium to large effects ($f = 0.3$) would require 191 participants. In total, 201 people participated in our research. Since the chatbot conversation's language was English, we targeted participants from English-speaking countries (US, Canada, UK) and asked them additionally to self-report their English proficiency. We excluded four participants who reported intermediate and advanced proficiency and only retained participants fluent in English (11) or native speakers (186). The concluding sample consisted of 197 participants: 94 identifying as male, 93 identifying as female, seven identifying as non-binary or third gender, and three choosing not to disclose their gender. Their age spanned from 19 to 54 with an average age of 34.42 ($SD = 10.30$). Around 30% of our participants stated having a high school diploma or equivalent, 45% a bachelor's degree, 12% an associate's degree, and 10% a master's degree. An additional four participants held professional degrees, and two had a doctorate. Our sample rated their affinity for technology interaction (ATI) as fairly high, with an average of 4.81 ($SD = 0.84$) (on a 7-point Likert scale, max seven). This tendency is also reflected in participants' reported experience with chatbots ($M = 4.57$, $SD = 1.43$, 7-point Likert scale, max seven). In contrast, our participants stated only to have little astronomy knowledge ($M = 2.51$, $SD = 1.42$, 7-point Likert scale, max seven).

## 4.5 Data Processing

We performed an outlier analysis calculating the Mahalanobis Distance [44] for our target variable trust and identified three participants with a probability of less than .001, which we removed from our dataset (final sample size of $N = 194$). After removing outliers, the group sizes ranged from 46 to 51 people per condition, resulting in an almost equal distribution of participants across stimuli conditions.

All our questionnaires use Likert scales, which produce ordinal data. However, the data can be assumed interval scaled if certain precautions are applied, such as (1) using a high-point scale (seven

in our case), (2) only presenting participants with textual interpretations for the min and max value (strongly disagree to strongly agree), and (3) averaging scales across multiple items [52]. To follow precaution (3), we created sum values for each questionnaire. For that purpose, we first calculated Cronbach's Alpha for the scales trust ($N$ = 197, four items, $\alpha$ = .881), subjective transparency ($N$ = 197, three items, $\alpha$ = .920), and social presence ($N$ = 197, five items, $\alpha$ = .952). All resulting alpha levels were very high, meaning the items of each questionnaire appear to be measuring the same construct; thus, we can aggregate the questionnaires' items and use the average across all items as a cumulative value for each scale. Since, in our case, all three assumptions described above have been met, we can consider our data interval scaled.

We assessed our data for normality with Shapiro-Wilk tests, finding that trust, social presence, subjective transparency, and response speed did not follow a normal distribution. Levene's test indicated no homogeneity of variance (.033) for trust and response speed but not for social presence and subjective transparency. Given our large and equally sized groups, t-tests and ANOVAs remain reliable despite these deviations [77].

## 4.6 Results

*4.6.1 Manipulation Check.* To assess the effectiveness of our manipulation of the chatbot's response speed, we analysed participants' rating of perceived response speed (item "I felt the response time for the chatbot to answer my question is...", assessed on a 7-point Likert scale from 1 = "very slow" to 7= "very fast"). An independent sample t-test showed that **participants exposed to the instant delay condition perceived the chatbot as responding significantly faster** ($M$ = 6.43, $SD$ = 1.15) as compared to those in the dynamic delay condition ($M$ = 5.65, $SD$ = 1.27; $t(189.93) = -4.576$, $p < .001$, Cohen's d of .658[13]). While we can consider our manipulation successful, it also has to be noted that both conditions were rated as relatively fast. Figure 5a presents a visualization depicting variations in perceived response speed across the two response delay conditions.

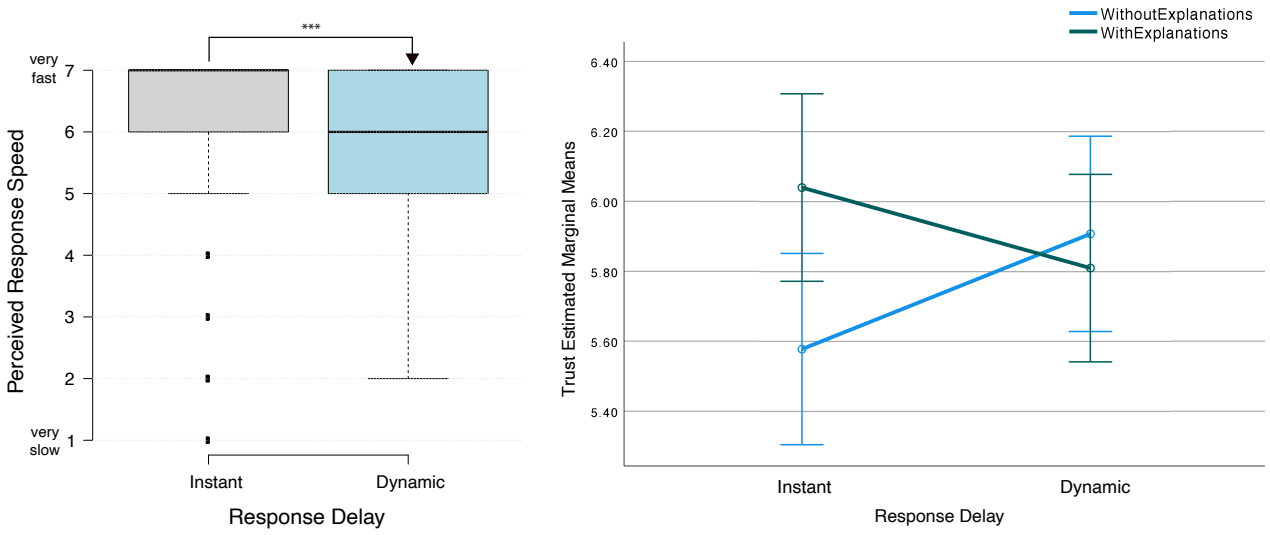
*4.6.2 Effect of Response Delay on Social Presence, Subjective Transparency, and Trust.* When considering all four conditions (with and without justifications), our participants reported the social presence of the instant response delay chatbot conditions on average at 2.75 ($SD$ = 1.53, 7-point Likert scale from 1 indicating low social presence to 7 indicating high social presence) and for the dynamic delay conditions minimally lower at 2.67 ($SD$ = 1.40), see Figures 6a. An independent sample t-test showed **no significant difference among the instant and delayed response conditions on how users perceived the chatbot in its social presence** ($t(1, 192) = -.282$, $p$ = .389). Similarly, another independent sample t-test showed **no difference between instant and dynamic response delay on subjective transparency** of the chatbot ($t(1, 192) = -.417$, $p$ = .339). For the instant delay conditions, subjective transparency was rated on average at 3.63 ($SD$ = 1.79) and marginally lower for the dynamic delay condition at 3.52 ($SD$ = 1.83), see Figure 6b. Lastly, no effect was found on users' trust in the chatbot between the instant and dynamic delay condition ($t(1, 192)$ = .309, $p$ = .379), only within the instant condition, see Figure 6c. When looking only at the two control conditions that did not have justifications, an independent sample t-test revealed a significant difference in trust between the instant and dynamic condition ($t(1, 81)$ = 1.725, $p$ = .044, Cohen's d of .353), i.e. the trust in the dynamic (no justification) condition was higher ($M$ = 5.91, $SD$ = 0.719) than the trust in the instant (no justification) condition ($M$ = 5.58, $SD$ = 1.10). No effects are observed when directly comparing the two conditions that include justifications.

*4.6.3 Effect of Justifications on Social Presence, Subjective Transparency, and Trust.* To investigate whether the justifications affected our dependent variables presence, transparency, and trust, we again conducted independent sample t-tests. We found no significant differences in the social presence ($t(1, 188) = -.439$, $p$ = .331) and on trust($t(1, 192)$ = 1.338, $p$ = .091). However, we found a significant difference in subjective transparency ($t(1, 192) = -2.095$, $p$ = .019, Cohen's d of .301). **Participants rated the transparency in the justification conditions (instant and dynamic) significantly higher (**$M$ = 3.84, $SD$ = 1.89**) compared to the condition without justifications (**$M$ = 3.30, $SD$ = 1.68**).** When looking into the individual items of the scale (see Table 3 in Appendix) using independent sample t-tests, we see significant differences in the items T2 ("*The chatbot is capable at addressing my issues*", $t(1, 192) = -1.796$, $p$ = .037, Cohen's d of .258), ST1 ("*I can access a great deal of information that explains how the system works*", $t(1, 192) = -2.182$, $p$ = .015, Cohen's d of .314) and ST3 ("*I felt that the amount of available information regarding the system's reasoning is large*", $t(1, 192) = -2.019$, $p$ = .022, Cohen's d of .290). For all three items, participants' rating was higher for the condition with a justification than without. Thus, we consider our **H1a** confirmed.

*4.6.4 Effect of Justifications on Mediating the Relationship between Response Delay and Trust.* To investigate a potential interaction effect between response delay and justifications, we conducted a two-way ANOVA. There was a statistically significant interaction between the effects of response delay and justification availability on trust ($F(1, 190)$ = 4.112, $p$ = .044, partial $\eta^2$ = .021[14], see Figure 5b). A follow-up simple main effects analysis showed that **trust was higher in the instant condition with justifications than without justifications** ($p$ = .018), but there was no effect of justifications in the dynamic condition ($p$ = .619). When looking at the descriptive data, we see that in the instant delay conditions, the trust rating was significantly higher when justifications were present ($M$ = 6.04, $SD$ = 0.95) than when they were not present ($M$ = 5.58, $SD$ = 1.10). In the dynamic response delay conditions, trust was rated minimally lower in the condition with justifications ($M$ = 5.81, $SD$ = 1.02) as compared to without justifications ($M$ = 5.91, $SD$ = 0.72). We consider **H1b** partially confirmed, as the mediating effect of justifications is only occurring in one of our conditions.

[13] For Cohen's d in t-tests, Correll et al. [11] define 0.2 as small effect, 0.5 as medium effect, and 0.8 as large effect.

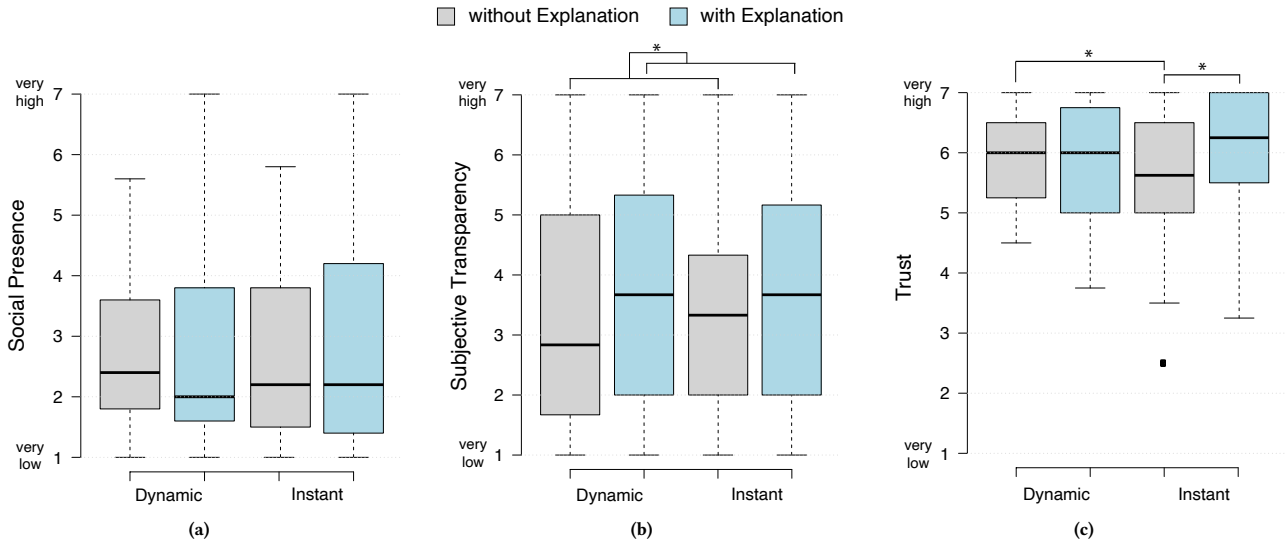
[14] Can be benchmarked against Cohen's ([10], pp. 278–280) criteria [58]: 0.2 as small effect, 0.5 as medium effect, and 0.8 as large effect.




(a) Manipulation Check

(b) Interaction Effect ResponseDelay*Justification on Trust

**Figure 5: Users perceived the response speed of the instant chatbot as significantly slower than in the dynamic condition (a) and the response delay, together with the justification, creates an interaction effect on users' perceived trust in the chatbot (b). The statistically significant difference is indicated by *** as $p < .001$.**




(a)

(b)

(c)

**Figure 6: Boxplots of our three dependent variables (a) social presence, (b) subjective transparency, and (c) trust, and how they are affected by our independent variables response delay and justification availability. The statistically significant differences are indicated by * as $p < .05$.**

*4.6.5 Other Effects.* During our analysis, we noticed that **chatbot experience is significantly correlated with the rating of subjective transparency**. While the Pearson correlation is also fairly robust toward violating the normality assumption, we here opted for a Spearman's correlation as it is recommended as better suited if we can not assume that the data has a linear relationship and if outliers are present [12]. The two variables are positively correlated ($r(194) = .239$, $p < .001$), and the analysis shows a moderate effect strength of .239.

Furthermore, a Spearman's Rho correlation test revealed **significant correlations between trust and social presence** (r(195)=.165, $p = .021$), **trust and subjective transparency** ($r(195) = .247$, $p < .001$) as well as **social presence and subjective transparency** ($r(195) = .216, p = .002$). Thus, the perception of all three constructs appears to be positively related in our study.

## 5 DISCUSSION

The main goal of our work was to explore the influence of response delay and justifications on users' trust, as well as observe effects on social presence and subjective transparency as related concepts. Based on our research model (Figure 1), we hypothesized that providing justifications about the response delay would improve chatbot transparency and increase users' trust for both settings (instant vs. dynamic response delay). We were able to confirm the positive effect of justifications on subjective transparency (**H1a**) but only see a positive mediating effect of justifications on trust for the instant response delay condition (**H1b**). We will discuss our findings and implications towards the design of trustworthy chatbots using response delay and justifications as social cues.

### 5.1 Response delay as a trust-enhancing design feature

Our statistical analysis showed no significant difference between the instant and delayed response conditions on how users perceived the chatbot in its social presence, subjective transparency, and trust. However, the positive correlation between social presence and trust indicates that users who perceived a greater social presence from the chatbot tended to trust it more [39, 75]. This highlights the need to enhance the sense of social presence in chatbot interactions [64]. Our results also showed a correlation between perceived response delay and user trust, indicating that users are more likely to trust the chatbot that was perceived to respond fast. However, previous studies indicate that dynamic response delay can significantly increase user trust in the chatbot compared to instant responses [9]. Considering these diverging results, more research is needed looking into comparing more fine-grained levels of response delay but also the manipulation of other design choices such as output presentation (cf. Kojima et al. [38]) to ultimately derive insights towards using response delay as a trust-enhancing design feature. An important aspect while designing response delay in chatbots is to improve users' perception of the response delay. Our analysis showed that participants exposed to the instant delay condition perceived the chatbot's responses as significantly faster than the dynamic response. To improve users' experience, design components such as images, text, progress bar, background color, or multimedia can be used to influence users' perceived waiting time (PWT) [43].

### 5.2 Justifications as a trust-enhancing design feature.

In our study, justifications were used to make the chatbot transparent in terms of its response generation process. More specifically, the retrieval-based chatbot justifies its ability to provide a nearly instant response by describing the process of retrieving the most appropriate response. Similarly, the generation-based chatbot justifies the response delay by describing the dynamic generation of a response. While our results indicate that explaining the delay, whether instant or dynamic, does not significantly influence trust levels, a statistically significant interaction effect was found between justifications and the response delay conditions, indicating that providing a justification can significantly increase the trust levels of a retrieval-based chatbot (instant responses). However, in the dynamic response group, justifications did not significantly affect trust. This could indicate that providing an instant response requires a justification (explanation), while the delay for dynamic responses can be self-explanatory. One explanation of this effect is the pervasiveness and current hype around generation-based chatbots such as ChatGPT, which can only provide delayed responses since the model needs to generate the response based on the user's input. Following the assumption that dynamic response delays can increase user trust [9, 29], users may perceive dynamic responses as the default setting of generation-based chatbots. Thus, justifications would be less crucial to reach a sufficient level of trust. On the other hand, receiving an immediate response (from a retrieval-based chatbot) may require further justification about the system's functionality, i.e., explaining why the response is so fast or how the system generates the response. Furthermore, our analysis shows an effect of justifications on subjective transparency but not directly on trust. Subjective transparency positively affected user trust, with participants expressing higher trust when the chatbot explained its responses. Our study replicates the positive correlation between subjective transparency and user trust, emphasizing the importance of designing transparent chatbot-based systems. In our study, justifications were used to explain the response delay, which is a narrow aspect of a human-chatbot interaction. Explaining other chatbot functionalities, including the quality of the response, can build a sense of transparency and likely lead to more trust in the chatbot. Focusing on the design of justifications, the goal of the pretest was to identify which justification type has a larger effect on social presence and subjective transparency. While the *First-person + Detail* justification ranked best for both delay conditions regarding these two facets, further research is needed to explore other design options for effective justifications in chatbot-based systems.

### 5.3 Is "one-design-fits-all" appropriate for trust-enhancing chatbots?

One of the main challenges while designing chatbot-based systems that use social cues is the variability of the mental models users build when interacting with a chatbot [53]. Mental models are *"cognitive constructs that guide the user's understanding of the system, and how to operate it"* [61]. In our study, we use measures related to how users perceive different chatbot aspects, including subjective measures for transparency, response delay, and social presence. We investigated how justifications and response delays can be used as social cues to enhance user's trust. The marginal interaction effect between delay type and justification on trust suggests a nuanced relationship. While it is evident that subjective transparency positively correlates to users' trust, more exploration is needed to investigate how to enhance subjective transparency using social cues. Chatbot designers need to consider the different mental models that

**Table 2: Overview of Spearman Correlations. Significant correlations with a p-value $< .05$ are marked by an $^*$. Spearman's Rho can take values between -1 and +1, where 0 indicates no association, +1 is a perfect positive, and -1 is a perfect negative association [25].**

| | Subjective Transparency | Social Presence | Trust | Experience with Chatbots |
|---|---|---|---|---|
| Subjective Transparency | - | $r(195) = .216, p = .002^*$ | $r(195) = .247, p < .001^*$ | $r(194) = .239, p < .001^*$ |
| Social Presence | $r(195) = .216, p = .002^*$ | - | $r(195) = .165, p = .021$ | |
| Trust | $r(195) = .247, p < .001^*$ | $r(195) = .165, p = .021$ | - | |
| Experience with Chatbots | $r(194) = .239, p < .001^*$ | | | - |

users can build during their interaction, as well as how different users may perceive the designed social cues. Furthermore, user expectations and their abilities to build mental models can change as users get more experienced and familiarized with chatbots. Such differences in user characteristics highlight the need to shift from a "one-design-fits-all" approach towards developing chatbots that can be adapted to individual user characteristics, preferences and contextual information [29]. Since chatbots are frequently designed to interact with different types and groups of users, chatbot features should be designed considering such user characteristics at a community level. Either based on a user profile or an individual level, adaptive learning could allow chatbots to learn from prior interactions with a user (group) and create more personalized responses and behaviour patterns. Participatory Design (PD) can be used to involve different stakeholders and potential users in the design process to design community-based AI systems [40]. While this work focussed on a broad sample and use case, testing response behaviour in a specific context can provide further insights into the target group's preferences.

## 6 LIMITATIONS AND FUTURE WORK

As appropriate for a study exploring a novel concept such as textual justifications for response delay, our experiment has several limitations: We tested a limited set of justifications with two chatbot types, leaving the generalizability to other combinations, delay lengths, or response presentation modes (such as word-by-word presentation) for future research. As described in the discussion section, a one-size-fits-all approach, in which one justification is presented for every response, will likely not hold up in realistic scenarios. Instead, we recommend that future work explores conversation-adaptive justifications to provide more transparency in the answer generation, for example, by considering input and response length and complexity, similar to delays in human communication. The focus of this study was, however, a first exploration of users' perceptions of justifications for response delay. Thus, adaptive justifications were out of the scope of this work.

Further, we used video recordings of chatbot interactions as material for our study instead of a live interaction. While this was necessary to control the content of the conversation, we see a need for investigating response delays and justifications in realistic situations. In live chatbot interactions, users have direct control over the conversation flow, which could influence their perception of agency, the thoroughness with which participants read and process the text output, and, as a consequence, the perceived chatbot responsiveness and capabilities. Active participation might also affect the user's emotional engagement and, hereby, their trust in the chatbot. While all conditions were affected equally by this experimental setup, we see a need for future research to consider a within-subject study design that allows participants to engage with chatbots in real time, offering a more interactive and authentic experience.

Lastly, we acknowledge that while we could show that our two response delay conditions were perceived as significantly different, they were both still perceived as relatively fast. This could explain the absence of general effects observed between instant and dynamic conditions on all dependent variables (trust, presence, and transparency). We recommend further research exploring a wider range of conditions.

## 7 CONCLUSION

The increasing use of chatbots in many domains has underlined the need for more transparency in and through the design of such AI technology. Our study investigated the use of justifications, a form of textual explanation, for response delays and their impact on users' perceptions of social presence, subjective transparency, and trust. Our results show that the relationship is complex. Firstly, while the response delay itself – be it instant or dynamic – did not affect the chatbot's perceived social presence or subjective transparency, there was an interesting effect of justifications. Specifically, justifications led to an enhanced sense of subjective transparency across both the instant and dynamic response delay conditions, showing the importance of providing justification for a chatbot's response speed. Moreover, our findings suggest an interaction between response delay and justifications concerning trust. While justifications increased trust for instant response delays, they had no significant impact on the dynamic delay condition. This underlines the complexity of this relationship and creates the need for further studies to understand users' impressions and experiences in the two response conditions. These results also emphasize the need for designers and developers to prioritize transparency for the answer generation process, especially when the chatbot's responses are immediate, and consider response speed not just a fixed factor but potentially a design choice.

## REFERENCES

[1] Ritu Agarwal and Mani Wadhwa. 2020. Review of state-of-the-art design techniques for chatbots. *SN Computer Science* 1, 5 (2020), 246. https://doi.org/10.1007/s42979-020-00255-3

[2] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15, 2 (2023), 1–4. https:

//doi.org/10.7759/cureus.35179
[3] Theo Araujo. 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 (2018), 183–189. https://doi.org/10.1016/j.chb.2018.03.051
[4] Adella Bhaskara, Michael Skinner, and Shayne Loft. 2020. Agent Transparency: A Review of Current Theory and Evidence. *IEEE Transactions on Human-Machine Systems* (2020), 1–10.
[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[6] Judee K Burgoon, Valerie Manusov, and Laura K Guerrero. 2021. *Nonverbal communication*. Routledge, London, UK. https://doi.org/10.4324/9781003095552
[7] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (2017), 25–35. https://doi.org/10.1145/3166054.3166058
[8] Xusen Cheng, Xiaoping Zhang, Jason Cohen, and Jian Mou. 2022. Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. *Information Processing & Management* 59, 3 (2022), 102940. https://doi.org/10.1016/j.ipm.2022.102940
[9] Karma Choedak. 2020. *The Effect of Chatbot Response Latency On Users' Trust*. Ph. D. Dissertation. University of Oklahoma.
[10] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.
[11] Joshua Correll, Christopher Mellinger, Gary H McClelland, and Charles M Judd. 2020. Avoid Cohen's 'small','medium', and 'large'for power analysis. *Trends in cognitive sciences* 24, 3 (2020), 200–207. https://doi.org/10.1016/j.tics.2019.12.009
[12] Joost CF De Winter, Samuel D Gosling, and Jeff Potter. 2016. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods* 21, 3 (2016), 273.
[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[14] Stephan Diederich, Alfred Benedikt Brendel, and Lutz M Kolbe. 2020. Designing anthropomorphic enterprise conversational agents. *Business & Information Systems Engineering* 62 (2020), 193–209. https://doi.org/10.1007/s12599-020-00639-y
[15] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*. ACM, New York, NY, US, 211–223. https://doi.org/10.1145/3172944.3172961
[16] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.
[17] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009
[18] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e7785. https://doi.org/10.2196/mental.7785
[19] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for Customer Service: User Experience and Motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3342775.3342784
[20] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150
[21] Boris Galitsky and Saveli Goldberg. 2019. *Explainable Machine Learning for Chatbots*. Springer International Publishing, Cham, 53–83. https://doi.org/10.1007/978-3-030-04299-8_3
[22] Dennis F Galletta, Raymond M Henry, Scott McCoy, and Peter Polak. 2006. When the wait isn't so bad: The interacting effects of website delay, familiarity, and breadth. *Information Systems Research* 17, 1 (2006), 20–37. https://doi.org/10.1287/isre.1050.0073
[23] David Gefen and Detmar W Straub. 1997. Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS quarterly* (1997), 389–400.
[24] David Gefen and Detmar W Straub. 2004. Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services. *Omega* 32, 6 (2004), 407–424. https://doi.org/10.1016/j.omega.2004.01.006
[25] Andrew R Gilpin. 1993. Table for conversion of Kendall's Tau to Spearman's Rho within the context of measures of magnitude of effect for meta-analysis. *Educational and psychological measurement* 53, 1 (1993), 87–92.
[26] Lorentsa Gkinko and Amany Elbanna. 2023. The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users. *International Journal of Information Management* 69 (2023), 102568. https://doi.org/10.1016/j.ijinfomgt.2022.102568
[27] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward Establishing Trust in Adaptive Agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (Gran Canaria, Spain) *(IUI '08)*. Association for Computing Machinery, New York, NY, USA, 227–236. https://doi.org/10.1145/1378773.1378804
[28] Ulrich Gnewuch, Stefan Morana, Marc TP Adam, and Alexander Maedche. 2018. Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *26th European Conference on Information Systems: Beyond Digitization-Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, June 23-28, 2018. Ed.: U. Frank*. 143975.
[29] Ulrich Gnewuch, Stefan Morana, Marc TP Adam, and Alexander Maedche. 2022. Opposing Effects of Response Time in Human–Chatbot Interaction: The Moderating Role of Prior Experience. *Business & Information Systems Engineering* 64, 6 (2022), 773–791. https://doi.org/10.1007/s12599-022-00755-x
[30] Christina N. Harrington and Lisa Egede. 2023. Trust, Comfort and Relatability: Understanding Black Older Adults' Perceptions of Chatbot Design for Health Information Seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 120, 18 pages. https://doi.org/10.1145/3544548.3580719
[31] Thomas Holtgraves and Tai-Lin Han. 2007. A procedure for studying online conversational processing using a chat bot. *Behavior research methods* 39, 1 (2007), 156–163. https://doi.org/10.3758/BF03192855
[32] Baptiste Jacquet, Jean Baratgin, and Frank Jamet. 2019. Cooperation in online conversations: the response times as a window into the cognition of language processing. *Frontiers in psychology* 10 (2019), 727. https://doi.org/10.3389/fpsyg.2019.00727
[33] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730
[34] Takeshi Kamita, Atsuko Matsumoto, Boyu Sun, and Tomoo Inoue. 2020. Promotion of Continuous Use of a Self-Guided Mental Healthcare System by a Chatbot. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 293–298. https://doi.org/10.1145/3406865.3418343
[35] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.
[36] Anjali Khurana, Parsa Alamzadeh, and Parmit K Chilana. 2021. ChatrEx: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, New York, NY, US, 1–11. https://doi.org/10.1109/VL/HCC51201.2021.9576440
[37] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. (1975).
[38] Hiroki Kojima, Dominique Chen, Mizuki Oka, and Takashi Ikegami. 2021. Analysis and Design of Social Presence in a Computer-Mediated Communication System. *Frontiers in psychology* 12 (2021), 641927. https://doi.org/10.3389/fpsyg.2021.641927
[39] Elisa Konya-Baumbach, Miriam Biller, and Sergej von Janda. 2023. Someone out there? A study on the social presence of anthropomorphized chatbots. *Computers in Human Behavior* 139 (2023), 107513. https://doi.org/10.1016/j.chb.2022.107513
[40] Yubo Kou and Xinning Gui. 2020. Mediating community-AI interaction through situated explanation: the case of AI-Led moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27. https://doi.org/10.1145/3415173
[41] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. ACM, New York, NY, US, 126–137. https://doi.org/10.1145/2678025.2701399
[42] Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems* 16, 10 (2015), 1.
[43] Younghwa Lee, Andrew NK Chen, and Virginia Ilie. 2012. Can online wait be managed? The effect of filler interfaces and presentation modes on perceived waiting time online. *Mis Quarterly* (2012), 365–394. https://doi.org/10.2307/41703460
[44] Christophe Leys, Olivier Klein, Yves Dominicy, and Christophe Ley. 2018. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of experimental social psychology* 74 (2018), 150–156.

[45] Abbas Saliimi Lokman and Mohamed Ariff Ameedeen. 2019. Modern chatbot systems: A technical review. In *Proceedings of the Future Technologies Conference (FTC) 2018: Volume 2*. Springer, Cham, 1012–1023. https://doi.org/10.1007/978-3-030-02683-7_75

[46] Bei Luo, Raymond YK Lau, Chunping Li, and Yain-Whar Si. 2022. A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 1 (2022), e1434.

[47] Joseph B Lyons, Izz aldin Hamdan, and Thy Q Vo. 2023. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior* 138 (2023), 107473.

[48] Sara Moussawi, Marios Koufaris, and Raquel Benbunan-Fich. 2021. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electronic Markets* 31 (2021), 343–364. https://doi.org/10.1007/s12525-020-00411-w

[49] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied* 7, 3 (2001), 171. https://doi.org/10.1037/1076-898X.7.3.171

[50] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153

[51] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, New York, NY, US, 72–78. https://doi.org/10.1145/191666.191703

[52] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education* 15 (2010), 625–632. https://doi.org/10.1007/s10459-010-9222-y

[53] Stine Ordemann. 2020. *"Understanding How Chatbots Work": An Exploratory Study of Mental Models in Customer Service Chatbots.* Master's thesis. University of Oslo.

[54] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[55] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 1, Article 140 (jan 2020), 67 pages.

[56] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* 151 (2021), 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

[57] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10, 10 (1996).

[58] John TE Richardson. 2011. Eta squared and partial eta squared as measures of effect size in educational research. *Educational research review* 6, 2 (2011), 135–147.

[59] Scott Schanke, Gordon Burtch, and Gautam Ray. 2021. Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research* 32, 3 (2021), 736–751. https://doi.org/10.1287/isre.2021.1015

[60] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, Arizona) *(AAAI'16)*. AAAI Press, 3776–3783.

[61] H Sharpe, Y Rogers, and J Preece. 2007. Interaction design: beyond human-computer interaction 2nd ed. *H Sharpe, Y Rogers, and J Preece* (2007).

[62] John Short, Ederyn Williams, and Bruce Christie. 1976. The social psychology of telecommunications. *Contemporary Sociology* 7, 1 (1976), 32–33.

[63] Joseph B Walther and Lisa C Tidwell. 1995. Nonverbal cues in computer-mediated communication, and the effect of chronemics on relational communication. *Journal of Organizational Computing and Electronic Commerce* 5, 4 (1995), 355–378.

[64] Yuping Wang, Wei-Chieh Fang, Julia Han, and Nian-Shing Chen. 2016. Exploring the affordances of WeChat for facilitating teaching, social and cognitive presence in semi-synchronous language exchange. *Australasian Journal of Educational Technology* 32, 4 (2016). https://doi.org/10.14742/ajet.2640

[65] Justin D Weisz, Mohit Jain, Narendra Nath Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: teaching strategies for successful human-agent interactions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 448–459. https://doi.org/10.1145/3301275.3302290

[66] Daricia Wilkinson, Öznur Alkan, Q Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P Knijnenburg, and Elizabeth Daly. 2021. Why or why not? The effect of justification styles on chatbot recommendations. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–21. https://doi.org/10.1145/3441715

[67] Philipp Wintersberger, Tobias Klotz, and Andreas Riener. 2020. Tell me more: Transparency and time-fillers to optimize chatbots' waiting time experience. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, New York, NY, US, 1–6. https://doi.org/10.1145/3419249.3420170

[68] Yu Wu, Zhoujun Li, Wei Wu, and Ming Zhou. 2018. Response selection with topic clues for retrieval-based chatbots. *Neurocomputing* 316 (2018), 251–261. https://doi.org/10.1016/j.neucom.2018.07.073

[69] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3506–3510. https://doi.org/10.1145/3025453.3025496

[70] Kun Xu, Mo Chen, and Leping You. 2023. The Hitchhiker's Guide to a Credible and Socially Present Robot: Two Meta-Analyses of the Power of Social Cues in Human–Robot Interaction. *International Journal of Social Robotics* 15, 2 (2023), 269–295. https://doi.org/10.1007/s12369-022-00961-3

[71] Wei Xu. 2019. Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *Interactions* 26, 4 (jun 2019), 42–46. https://doi.org/10.1145/3328485

[72] Chiahui Yen and Ming-Chang Chiang. 2021. Trust me, if you can: a study on the factors that influence consumers' purchase intention triggered by chatbots based on brain image evidence and self-reported assessments. *Behaviour & Information Technology* 40, 11 (2021), 1177–1194.

[73] Jennifer Zamora. 2017. I'm Sorry, Dave, I'm Afraid I Can't Do That: Chatbot Perception and Expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction* (Bielefeld, Germany) *(HAI '17)*. Association for Computing Machinery, New York, NY, USA, 253–260. https://doi.org/10.1145/3125739.3125766

[74] Ruijing Zhao, Izak Benbasat, and Hasan Cavusoglu. 2019. Do users always want to know more? Investigating the relationship between system transparency and users' trust in advice-giving systems. *In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden, June 8-14, 2019. Research-in-Progress Papers*. (2019).

[75] Naim Zierau, Korbinian Flock, Andreas Janson, Matthias Söllner, and Jan Marco Leimeister. 2021. The influence of AI-based chatbots and their design on users' trust and information sharing in online loan applications. In *Hawaii International Conference on System Sciences (HICSS).-Koloa (Hawaii), USA*. IEEE, New York, NY, US. https://doi.org/10.24251/HICSS.2021.666

[76] Naim Zierau, Michael Hausch, Olivia Bruhin, and Matthias Söllner. 2020. Towards Developing Trust-Supporting Design Features for AI-Based Chatbots in Customer Service.. In *ICIS*, Vol. 2020. 1–9.

[77] D.W. Zimmerman. 2014. Robust Statistical Tests. In *Encyclopedia of Quality of Life and Well-Being Research*, A.C. Michalos (Ed.). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_2529

# A MAIN STUDY QUESTIONNAIRE

The questionnaires deployed in the main study are presented in Table 3. For trust, social presence, subjective transparency, and affinity for technology interaction, we opted for validated questionnaires. For the chatbot experience and the manipulation check, we created new items based on similar existing questions used in prior work. The items for astronomy knowledge were specifically designed for the purpose of this study.

**Table 3: Items used in the main study questionnaire and their references. All questionnaires except for the manipulation check applied 7-point Likert scales from 1 = "strongly disagree"; 7 = "strongly agree").**

| Item | | Reference |
|---|---|---|
| | **Social Presence** | Gefen & Staub [23] |
| SP1 | I felt a sense of human contact with the chatbot. | |
| SP2 | I felt a sense of personalness with the chatbot. | |
| SP3 | I felt a sense of sociability with the chatbot. | |
| SP4 | I felt a sense of human warmth with the chatbot. | |
| SP5 | I felt a sense of human sensitivity with the chatbot. | |
| **Subjective Transparency** | | Zhao et al. [74] |
| ST1 | I can access a great deal of information that explains how the system works. | |
| ST2 | I can see plenty of information about the system's inner logic. | |
| ST3 | I felt that the amount of available information regarding the system's reasoning is large. | |
| **Trust** | | Cheng et al. [8] |
| T1 | I felt the chatbot is honest and truthful. | |
| T2 | I felt the chatbot is capable of addressing my issues. | |
| T3 | I felt the chatbot's behavior and response can meet my expectations. | |
| T4 | I trust the answers provided by chatbots. | |
| **Previous Chatbot Experience** | | Gnewuch et al. [29] |
| PCE1 | I am familiar with chatbot technologies. | |
| PCE2 | I use chatbots frequently. | |
| **Astronomy Knowledge** | | n.a. |
| AK1 | I have a lot to say regarding astronomy. | |
| AK2 | I know a lot about astronomy. | |
| **ATI – Affinity for Technology Interaction** | | Franke et al. [20] |
| ATI1 | I like to occupy myself in greater detail with technical systems. | |
| ATI2 | I like testing the functions of new technical systems. | |
| ATI3 | I predominantly deal with technical systems because I have to. | |
| ATI4 | When I have a new technical system in front of me, I try it out intensively. | |
| ATI5 | I enjoy spending time becoming acquainted with a new technical system. | |
| ATI6 | It is enough for me that a technical system works; I don't care how or why. | |
| ATI7 | I try to understand how a technical system exactly works. | |
| ATI8 | It is enough for me to know the basic functions of a technical system. | |
| ATI9 | I try to make full use of the capabilities of a technical system. | |
| **Manipulation Check** (7-point Likert scale, 1 = "very slow"; 7 = "very fast") | | Galletta et al. [22] |
| MC1 | I felt the response time for the chatbot to answer my question is... | |