“Can freedom lead to trust & satisfy?” Mixed research: How does the user’s autonomy in interaction influence the trust perception and user experience

Jiaqi Lai
BSc/MSci Psychology and Language Sciences
University College London
2024

Supervised by:
Dr Martin Dechant

Word Count:
7497

Abstract

Trust has been identified as a core aspect of the success of human-ai interaction. Yet aspects which contribute to this core aspect remain unexplored but can have a critical influence. One of these aspects is the user’s experience of autonomy. Therefore, this study aimed to explore the impact of human autonomy on users' perceptions of trust and satisfaction with conversational agents (CAs), comparing interactions between Large Language Model (LLM)-powered and traditional branching dialogue CAs. Findings revealed that while users interacting with LLM CAs experienced higher autonomy, this did not translate into increased trust or satisfaction. Additionally, trust was positively correlated with satisfaction, yet autonomy did not significantly influence this relationship. Various factors, including interaction sequence order, task workload, and user familiarity, might have contributed to these outcomes, alongside technical issues with LLM CAs and a preference for the CA's problem-solving abilities over autonomy. Future research should employ Confirmatory Factor Analysis (CFA) and Structural Equation Modelling (SEM) to refine the measurement of trust and satisfaction and examine the mediation effect between autonomy, trust, and satisfaction. Our work contributes towards a better understanding about core requirements for the development of user’s trust in conversational agents.

1. Introduction

In the field of artificial intelligence (AI), Large Language Models (LLMs) represent a pivotal advancement in human-computer interaction (HCI), because of its application and potential (Kaddour et al., 2023). The introduction of GPT-3 by OpenAI has not only garnered widespread attention but has also broadened the application of LLMs beyond the traditional boundaries of computer science, enabling new ways to interact with computers and AI by natural languages, highlighting the versatility and potential of these models in enhancing user experiences (Radford et al., 2019).

Historically, the natural language interaction between humans and computers has always been a goal of interaction studies (Ogden & Bernick, 1997; Manaris, 1998). The evolution of LLMs has improved this dynamic, offering an unprecedented level of understanding and flexibility, thereby revolutionising the way we interact with digital systems. This shift has significant implications for the design and implementation of conversational agents (CAs) or chatbots, which have become increasingly prevalent in consumer-facing applications in recent years, offering a myriad of interactive tasks through voice or text interfaces. The implementation of CAs, exemplified by their use in the Dutch airline KLM in 2016, has been shown to enhance operational efficiency by 20% within communication departments (Zemčík, 2019), underscoring the significant potential of CAs in interaction design (Grand View Research, 2021).

Despite these advancements, the adoption and integration of AI-infused systems into human societies hinge on a crucial factor: trust. The acceptance of these technologies by users as well as developers is predicated on their trustworthiness and the perceived value they bring to human-computer interaction (Glikson & Woolley, 2020). This research seeks to examine the differential impacts of interaction design on user experience, with a focus on the novel interaction paradigm introduced by LLMs, in the context of maintaining consistent performance levels. It is within this context that this study aims to explore two core questions:

Q1: Whether LLMs, through natural language processing, can indeed enhance human autonomy in interaction compared to traditional CA interaction form.

Q2: Whether this heightened autonomy can lead to increased user perception of trust and satisfaction.

2. Related works

Traditional branching dialogue tree

Among different methods, the traditional branching dialogue tree stands out for its historical significance and widespread application in various domains, including game design and early commercial interfaces (Fraser et al., 2018). The branching dialogue tree, as elucidated by Ellison (2008), originated primarily within the gaming industry, offering players a structured way of navigating conversations by selecting from multiple pre-scripted options. It also has many commercial applications in many companies' phone call automatic responding systems and primary website guidance. The commercial CA with a dialogue tree usually provides several options and the user could choose among them to continue the interaction. They are closer to the menu-based design compared to the gaming application, allowing users the flexibility to revisit previous choices and alter their decisions (Bailly et al., 2016; Paap & Cooke, 1997). Adopting the branching dialogue tree in this experiment is a control condition with a minimal autonomy index for comparative analysis against Large Language Models (LLMs).

The concept of human autonomy and its importance

The concept of autonomy was traditionally considered a philosophical study. Early empiricist philosophers defined the autonomy as unqualified freedom of choice (Bentham, 1789; Mill, 1861). To be more specified in the modern HCI research content, Lammers (2016) delineated autonomy as the capacity to act independently, without being influenced by others, and in accordance with one's desires. In our study, we adopt the terminology from Calvo et al. (2020) to describe this person-centric notion as "human autonomy." Conversely, we refer to the ability of non-human agents or machines to operate autonomously, without human

intervention, as "machine autonomy." While existing literature predominantly focuses on the implications of machine autonomy from an interaction perspective (Rödel et al., 2014; Lucia-Palacios & Pérez-López, 2023; Ulfert et al., 2022), research on human autonomy and its effects on user satisfaction is comparatively scarce. Calvo et al. (2020) argued that the predominant focus on machine autonomy is insufficient, and for HCI studies and AI development to truly benefit humanity, investigating human autonomy is essential. Meanwhile, Sankaran et al, (2020) argued that the conflict between human autonomy and machine autonomy would be more intense, the loss of human autonomy was the primary concern toward AI development (Anderson et al., 2018). Thus, respecting human autonomy would be one of the primary principles in the HCI development (EU, 2021), and the bond toward the trust of AI (Sankaran et al., 2020).

Self-determination theory

Self-determination theory (SDT), as formulated by Ryan and Deci (2017), serves as a foundational theoretical framework for our experiment, explaining why the increase in users' human autonomy should be correlated with an increase in their perception of satisfaction and trust. SDT delves into the dynamics that catalyse human actions, focusing particularly on the extent to which behaviours are autonomously initiated and sustained. SDT articulates that autonomy is one of the three fundamental psychological needs deemed essential for the promotion of healthy psychological development, optimal functioning, and the energisation of human activity. Following the structure of SDT, autonomy can be conceptualised as being able to act aligned with one's objectives and fulfilling volition over the objectives that have been enacted (Deci & Ryan, 2000).

Further application of SDT within HCI and technology acceptance, as examined by Tyack and Mekler (2020), and Roca and Gagné (2008), demonstrates SDT's relevance in evaluating user experiences and technology use. Specifically, an SDT-enhanced Technology Acceptance Model (TAM) reveals that fulfilling users' autonomy and competence needs significantly impacts their continued IT use, mediated by increased perceived usefulness and enjoyment. This body of research underscores the utility of SDT in understanding and designing for optimal user

engagement and technology adoption.

The concept of trust and its importance

According to Lee & See (2004), trust has emerged as a crucial mechanism for navigating the increasing intricacy found within technological advancements, organisational structures, and the spectrum of human interactions. It plays a pivotal role in the domain of HCI, which significantly influences both the adoption of technology and the overall user experience (Ha & Stoel, 2009). Rousseau et al. (1998) suggest that at its core, trust encapsulates the notion of positive expectations and the willingness to be vulnerable. Consequently, this research adopts the definition posited by Mayer et al. (1995), understanding trust as the confidence of one entity (the trustor) in making itself vulnerable to another (the trustee), with the expectation that the trustee will execute an action critical to the trustor, regardless of the trustor's capacity to oversee or control the trustee.

Initial trust was defined as the trust formed immediately following a user's first interaction with a system (Wang & Benbasat, 2005). This focus stems from the understanding that users depend on initial trust to gauge the likelihood of future interactions with systems with which they are unfamiliar. In such scenarios, users' perceptions of uncertainty and risk regarding system use become markedly pronounced (McKnight et al., 2002), serving as a benchmark in their trust perception towards that system. Similarly, Conversational Agents (CAs) powered by Large Language Models (LLMs) represent a relatively novel mode of interaction for both the market and users. Thus, it is imperative to begin our study with users' first impressions of the system. Guided by this theoretical rationale, our experimental design intentionally employs a simulated background to situate the interaction within the context of the users' first use of the system.

Research Hypothesis

Understanding the subject experience is a pivotal aspect of this study, with the initial hypothesis aimed at examining the effects of the interaction task sequence on participants,

thereby mitigating its potential impact on data interpretation. Prior research, including Cockburn et al. (2017), has highlighted the significant role that interaction sequence plays in shaping hedonic evaluations, while Kahneman et al. (1993) and Shteingart et al. (2013) underscored its influence on memory and overall experience. To effectively manage this variable, the study utilised a randomisation process to allocate participants into two distinct groups, facilitating a between-subjects comparison to investigate the sequencing effects. This methodological approach not only underscores the rigorous planning of the experiment but also establishes a foundational framework for exploring the underlying hypothesis.

Hypothesis 1:

The participant's interaction sequence of interaction tasks will lead to a significant difference in the participant's perception of autonomy, trust, and satisfaction.

The next step is to investigate the design validity of variable control. Wei et al. (2022) investigated the topic of the "emergent ability" in LLMs, pointing out that the LLMs could have a significant increase in language processing performance compared to the smaller models. Meanwhile, Webb et al. (2023) argued that the LLMs like GPT-3, had a "surprisingly strong capacity" for abstract pattern induction. The emergent ability from the GPT-3 is capable of finding zero-shot solutions to a wide range of analogy problems. Such paradigm shifts in interaction design, are enabled by advancements in the performance of natural language processing. The stronger flexibility, language processing ability and abstract pattern induction of a conversational agent could enable the agent to make valid responses far beyond the pre-scripted content, and better deal with ambiguous content sorting issues (Miner et al., 2016). As a result, the interaction restriction in LLM conversations would be less for the user. Thus, we predict that the LLM-powered CA could lead the user to have a higher perceived human autonomy than the traditional branching dialogue CA (TRA CA).

Hypothesis 2:

The participant's autonomy status in the LLM CA condition is higher than the TRA CA condition.

Perceived trust, satisfaction & interaction of autonomy

In exploring the relationship between user autonomy and trust perception in human-computer interactions, this study draws upon foundational principles from Social Exchange Theory (SET) to probe how interaction quality significantly influences users' trust and subsequent technology engagement (Nasirian, F., Ahmadian, M., & Lee, O. K. D., 2017). The synthesis of communication methods that are closer to human-to-human dialogue, particularly through the lens of anthropomorphism, suggests a positive correlation with trust perception. This notion is supported by Yen & Chiang (2021), who employed both self-reported and neuroimaging approaches to substantiate the positive impact of anthropomorphic features on trust in conversational agents. The capability of LLMs to enable interaction by natural language processing and generate corresponding natural responses demonstrates a higher social presence and provides a closer user experience toward human communications.

Furthermore, this investigation considers the degree of user autonomy in information input and choice as a pivotal element for fostering trust. Initial explorations into conversational agents reveal a transition from rule-based to more flexible, user-driven interactions. However, Taylor-Giles (2014) critiques the early stages of conversational agent development for their limited response capability and pre-scripted dialogues, which inadvertently undermine user agency and experience. Echoing these sentiments, Ariely (2000) highlights how user experiences are shaped by the level of control over the information exchanged with the system. Traditional branching dialogue systems, despite their advanced anthropomorphic responses, inherently restrict user choices to predetermined options, engendering a pronounced sense of limitation and consequently diminishing trust and satisfaction.

Combined the factors above, this study led to the following hypothesis.

Hypothesis 3:

Higher autonomy status of participants results in higher perceived trust and satisfaction.

From the perspective of HCI, the relationship between user trust and satisfaction was

discussed a lot. Balaji (2019) conducted an experimental investigation utilising factor analysis, positioning trust as a crucial indicator of satisfaction. In contrast, Apraiz et al. (2023) highlighted that user-perceived trust and satisfaction should be regarded as distinct, emotion-related factors in interaction studies, despite their close relationship. Despite this distinction, the linkage between trust and satisfaction cannot be understated, as evidenced by Lankton et al. (2014). Their research demonstrated a direct influence of user trust on satisfaction within systems. However, this result's generalisability to the present study was limited because they focused on information systems with fewer human-like features. Building on this foundation, Mostafa & Kasamani (2022) delved into the focus on AI chatbots and justified that the initial trust could significantly increase the user's intention to use, which is also been recognised had a strong positive correlation with satisfaction (Bokhari, 2005).

According to these materials, the influence of human autonomy was coexistent between both trust and satisfaction. Considering these considerations, the concept of human autonomy emerges as a potential interactive effect in the trust-satisfaction relationship. This study hypothesised the role of autonomy implies that a sense of freedom and control in their own interaction could amplify the positive effects of trust on satisfaction. This conjecture posits that users who trust a CA and perceive a degree of autonomy in their interactions are likely to report higher satisfaction levels compared to those who feel their choices are limited. Given the lack of empirical research directly addressing the interplay among these three concepts, this study aims to investigate the hypothesis that autonomy can enhance the relationship between trust and satisfaction in the context of HCI.

Hypothesis 4:

The participant’s autonomy status is a significant interaction of the relationship between the participant’s perception of trust and attractiveness.

3. Methodology

3.1 Material

The experiment was presented on the participants' devices (mainly the laptop); Gorilla Experiment Builder was used to build, conduct the online experiment and collect the data. The interaction task of LLM CA is called the API of ChatGPT (gpt-4-1106-preview) from OpenAI. The JAVA Script and HTML code that implemented this calling was provided by Yuxiang Hua (University of Wollongong, Australia). The data analysis and visualisation used R language (version 4.3.1) in RStudio (2023.6.0.421). The software used in the thematic analysis is NVivo. The TRA CA was built by Gorilla Task Builder 2, structured by the traditional branching dialogue tree and all content and options were pre-generated by GPT-4. Figure.1 demonstrates the interface page of two interaction tasks.

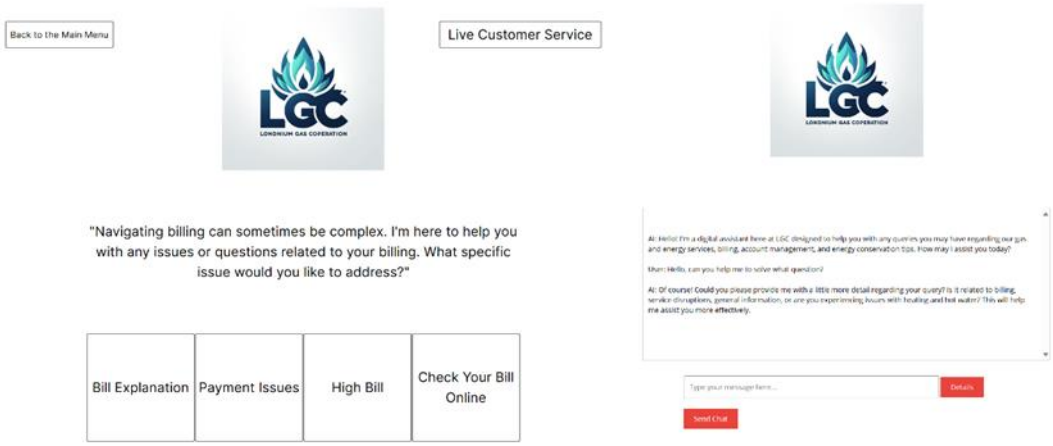


Figure.1 The Left part is the interface for TRA CA, the right part is for LLM CA

The background information and simulated challenge (attachment.1) were fictionally designed for the experiment. To conduct the variable control of the content quality of both CAs, all the pre-scripted content used in the TRA CA was generated by the same LLM as the one called in the LLM CA.

TRA CA

In the interaction task with the TRA CA, the user can choose different option buttons provided by the system, the user could back to the main menu if they want to change their options.

LLM CA

In another interaction task, the LLM CA will automatically greet the user and the user needs to use their own words to communicate with the CA by typing sentences.

3.2 Design

A between-subject design divided participants into 2 groups (Group A vs. Group B). Meanwhile, in each group, there was also a within-subject manipulation: participants in both groups needed to finish two interaction tasks that represented two CA-type conditions (LLM CA vs. TRA CA). The H1 is a priori hypothesis that has been used to ensure better variable control of the experiment to reduce the potential influence of the interaction sequence of tasks.

3.3 Measurements

After each interaction task in the Interaction phase, participants need to answer four sets of questions (attachment.2): NASA TLX, UEQ, self-rating question collection, and Open-ended question collection. In the Preference phase, participants need to finish two sets of questions (attachment.3): preference question collection, and demographic questionnaire.

3.3.1 Quantitative measurement

The NASA Task Load Index (NASA TLX)

The NASA TLX questionnaire used in this experiment includes 6 slider rating questions that measure the participant's subjective, each question measuring one subscale of the workload index of an interaction task (Hart & Staveland, 1988).

The User Experiment Questionnaire (UEQ)

The UEQ used in this experiment includes 26 slider rating questions, which can be translated into 6 subscales of user experience by the official analysis tool. According to the structural design principles of the User Experience Questionnaire (UEQ) articulated by Laugwitz et al. (2008) and Rauschenberger et al. (2012), the "attractiveness" subscale is formulated to capture the user's general impression of the experience, encompassing both pragmatic and

hedonic qualities. Therefore, in subsequent analyses, the attractiveness score will serve as a proxy for overall participant satisfaction.

The self-rating question collection

This collection contains 4 slider-rating questions that directly measure participants' subjective perception of the concept of "trust" and 2 slider-rating questions for the concept of "autonomy". In our experimental design, we introduced two distinct indicators, autonomy.1 and autonomy.2, to measure participants' subjective perceptions of autonomy. These indicators were operationalised through questionnaires addressing closely related yet distinct concepts: "having the freedom to do something" and "having control over the conversations in which one is engaged." Drawing from Lammer et al. (2016), we define "autonomy" as the capacity to remain uninfluenced by others, measured by autonomy.1, and "influence" as the capacity to exert control over others, assessed by autonomy.2. This differentiation not only aims to measure these constructs with greater precision but also to enable participants to clearly distinguish between concepts that are often conflated under the broad notion of "power." Based on the literature, we posit that a successful operationalisation of autonomy can be inferred if the data exhibit a significant variance in autonomy.1 scores across different CA-type conditions without a corresponding variance in autonomy.2 scores.

3.3.2 Qualitative measurement

The Open-ended question collection

This collection includes 3 open-ended questions that encourage participants to elaborate on their detailed autonomy perception and emotional expression in the interaction task.

The preference question collection

This collection has 1 slider rating question and 4 open-ended questions to investigate which CA they prefer to interact with and the reason behind their choice. The demographic

questionnaire includes 4 multiple-choice questions that collect participants' age, gender, education level and the field of their job/study.

3.4 Participant

A total number of 86 participants (Mean age =26.52, SD=10.40; Female = 69, Male = 15, non-binary/prefer not to say = 2) participated in the experiment through two different online experiment platforms (SONA = 56, Prolific = 30). 40 participants have finished undergraduate education (46.51%), and only 2 participants working or studying in the technology domain. There were no exclusion criteria for participant selection, so the distribution of the entire population could be examined. 7 participant who finished their experiment unusually fast (timespan < 960 seconds) or unusually slow (timespan > 3600 seconds) were excluded as outliers. All SONA-based participants were UCL psychology-related undergraduate students. To reduce the single resource influence of participants and increase the generalisability of the experiment results, this study introduced 30 Prolific-based participants who were randomly selected in the UK. This study was given ethics approval by the UCL Psychology & Language Sciences Dept.

3.5 Procedure

This experiment had three phases: Preparation, Interaction, and Preference. In the Preparation phase, participants needed to read the background information and the simulated challenge within it. In the interaction phase, participants will be randomly divided into two groups (group A & group B). In both groups, participants will experience two interaction tasks and answer a series of questionnaires after each interaction. The only difference between groups is the sequence of interaction tasks; group A will first interact with the LLM CA and then TRA CA, and vice versa in Group B. The questionnaires after each interaction were the same. In the Preference phase, participants will be ungrouped, they need to answer a questionnaire of their preferences of which CA and 4 followed open-ended questions, lastly their demographic information will be collected.

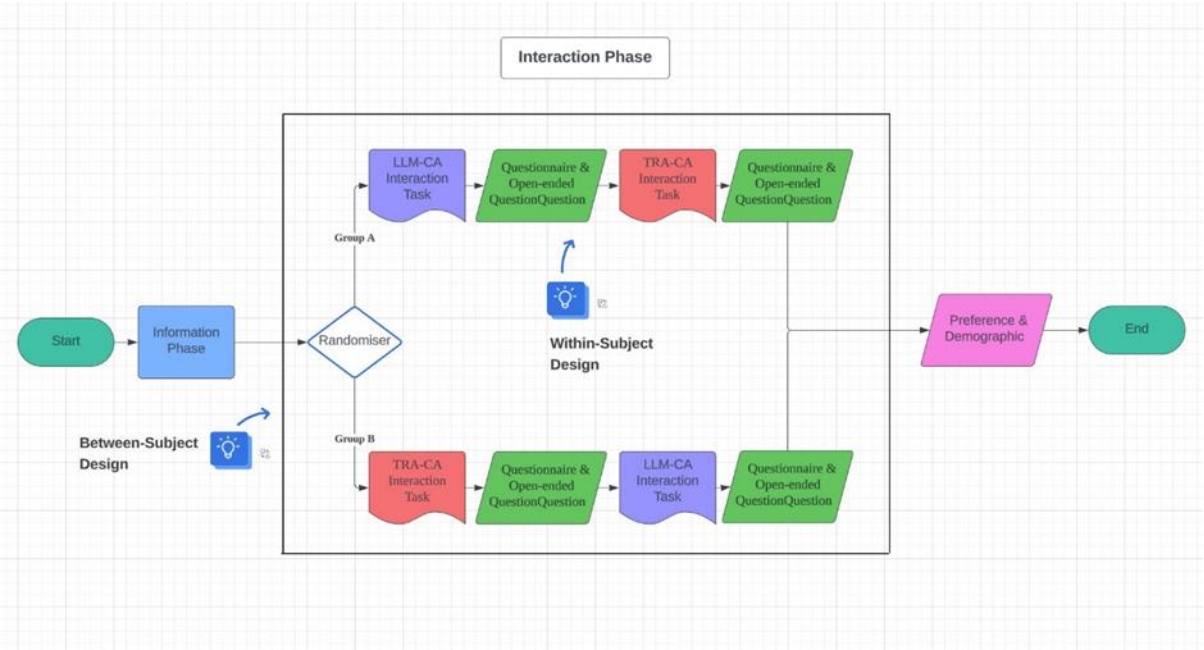


Figure.2 The flow chart of the experiment procedures

3.6 Analysis

H1: The dependent variable (DV) is the values of all quantitative questionnaires (NASA TLX, UEQ, self-rating question collection), and the independent variable (IV) is the participant grouping at the Interaction phase (2 levels: Group A vs. Group B). Wilcoxon rank sum tests were used to compare the grouping due to the non-normal distribution.

H2: The DV is the participant's rating score of a question “autonomy.1”, and the IV is the CA-type condition (2 levels: LLM CA vs. TRA CA) of the participants’ interaction task. Wilcoxon rank sum tests were used to do the comparison between CA-type conditions.

H3: DVs are the average perceived trust score of the participant, the 6 subscales of the UEQ, and the IV is the CA-type condition (2 levels: LLM CA vs. TRA CA). Wilcoxon rank sum tests and Wilcoxon signed-rank tests were used to do the comparison between CA-type conditions.

H4: The DV is the attractiveness from the UEQ as the representative of general satisfaction, the IV is the average perceived trust score, and the Moderating Variable is the CA-type condition. A linear regression model with interaction was used.

Qualitative data analysis employed thematic analysis, where responses underwent at least

two readings for the creation of codes, encapsulating both the explicit and underlying meanings expressed by participants. A bottom-up approach facilitated the gradual amalgamation of codes into themes, with specific quotes linked to each code and the most illustrative quote selected for association with the respective themes. The interconnections between potential themes and the coded data were meticulously examined during the theme review process. Themes received titles that most accurately and representatively summarised them. The coding process was completed independently by the researchers, who maintained a stance of conscious reflexivity throughout.

4. Result

4.1 Quantitative

Preparation

In the data preparation phase of our experiment, we employed the Shapiro-Wilk test to evaluate the normality of all relevant data, including autonomy, trust, NASA TLX, and UEQ. For example, the Shapiro-Wilk test for the trust.1 ($W = 0.96$, $p < .005$), NASA TLX.2 ($W = 0.8286248$, $p < .005$). The test results revealed a significant departure from normality across most data. These results indicate a non-normal distribution for all measures except Dependability ($p = 0.063$) and Efficiency ($p = 0.075$), which were only slightly higher than the threshold of $p = .005$. Given these findings of widespread non-normality, we opted for non-parametric methods for subsequent analyses.

H1: Influence of interaction sequence (Group A vs. Group B)

To justify the H1, we utilised the Wilcoxon rank sum test with continuity correction to investigate whether the sequence of interaction (group AB) led to any significant differences in measures such as trust, autonomy, NASA TLX, and UEQ components.

Trust.1 showed a significant difference between groups ($W = 4104.5$, $p < .005$), Figure.3 illustrate that Group A had a significantly higher average score than Group B. However, other

trust measures, such as trust.2 (W = 3206, p = 0.70), trust.3 (W = 3157.5, p = 0.83), and trust.4 (W = 3447, p = 0.22) did not demonstrate significant differences, suggesting a nuanced effect of the interaction sequence on trust.

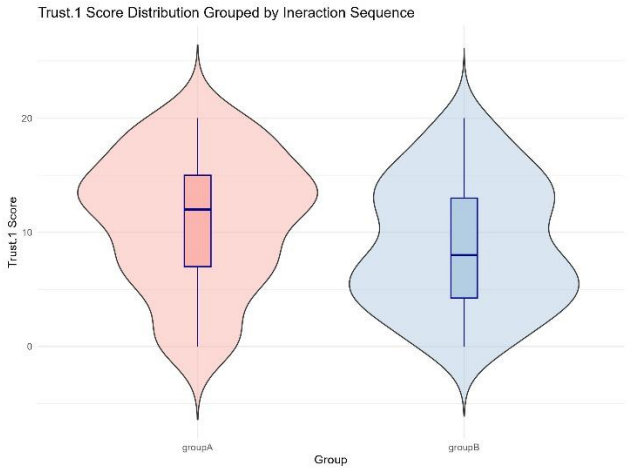


Figure.3 The comparison of trust.1 score distribution between the interaction sequence groups.

Similarly, autonomy measures (autonomy.1: W = 2959, p = 0.63; autonomy.2: W = 3557, p = 0.107) and preferences (W = 3318, p = 0.4383) showed no significant differences between groups, indicating that the sequence of interaction did not markedly affect these dimensions.

The NASA TLX components mostly did not show significant differences, except for NASA.TLX.4, which bordered on significance (W = 2536.5, p = 0.049), points to the sequence's potential impact on perceived workload. Other measures from UEQ, including attractiveness, perspicuity, dependability, stimulation, and novelty, also did not reveal significant differences, except for minor variations that suggest a potential, albeit not statistically significant, influence of the interaction sequence. In contrast, Efficiency displayed a significant group difference (W = 3993.5, p = 0.0017), highlighting its sensitivity to the interaction sequence.

H2: Relationship between autonomy and CA-type condition (LLM CA vs. TRA CA)

To conduct the relationship between the CA-type conditions and the participant perceived autonomy, this experiment used Wilcoxon rank sum tests to compare the differences between two CA-type conditions of two observed indicators (autonomy.1 and autonomy.2). In

experiment design, autonomy.1 has been used to measure the concept of “autonomy”, and a significant difference between CA-type conditions was observed (W = 4797.5, p < .005). Autonomy.2 was designed to measure the concept of “influence” and showed no significant difference between CA-type conditions (W = 3336.5, p = 0.45).

The result demonstrated a significant difference of measured value indicating the participant’s perception of “autonomy” between different CA-type conditions but not a significant difference of measured value indicating the participant’s perception of “influence”. Figure.4 illustrates that the autonomy score of the LLM CA condition (M = 11.41, SE = 0.62) was significantly higher than the TRA CA condition (M = 6.04, SE = 0.54), indicating the LLM CA condition could be the representative of the high status of participants’ perception of the concept of “autonomy”, and TRA CA condition represents the low status.

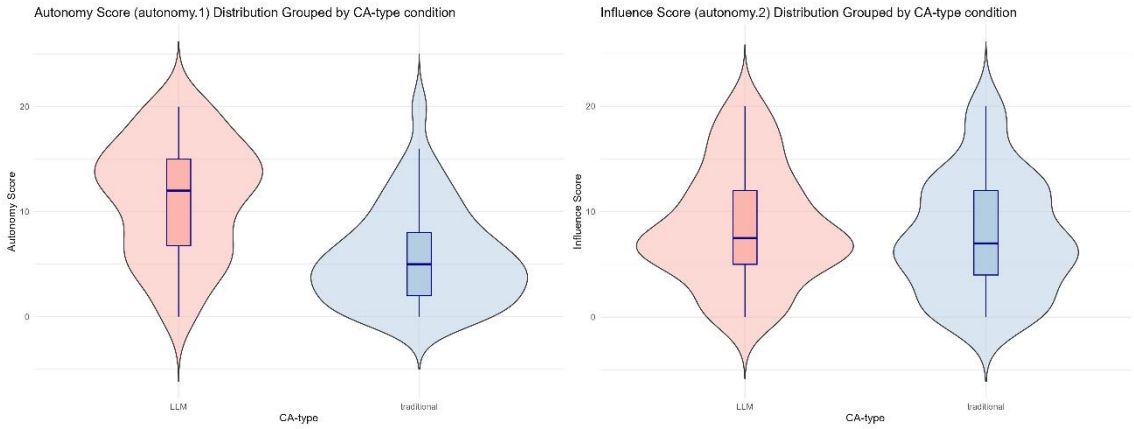


Figure.4 The comparison of autonomy score (autonomy.1) and influence score (autonomy.2) between two CA-type conditions.

H3: Relationship between autonomy and trust

The experiment first used the Wilcoxon rank sum tests revealing nuanced how two CA-type conditions influence the concept of "trust" across four indicators. The first trust indicator showed a very limited marginal significant difference (W = 2509.5, p = .033), and the subsequent indicator, trust.2 (W = 2655, p = .11), trust.3 (W = 2962.5, p = .58), and trust.4 (W = 2708, p = .15), demonstrated no significant differences. Figure.5 shows that the TRA CA

conditions had higher scores than the LLM CA conditions across all four indicators.

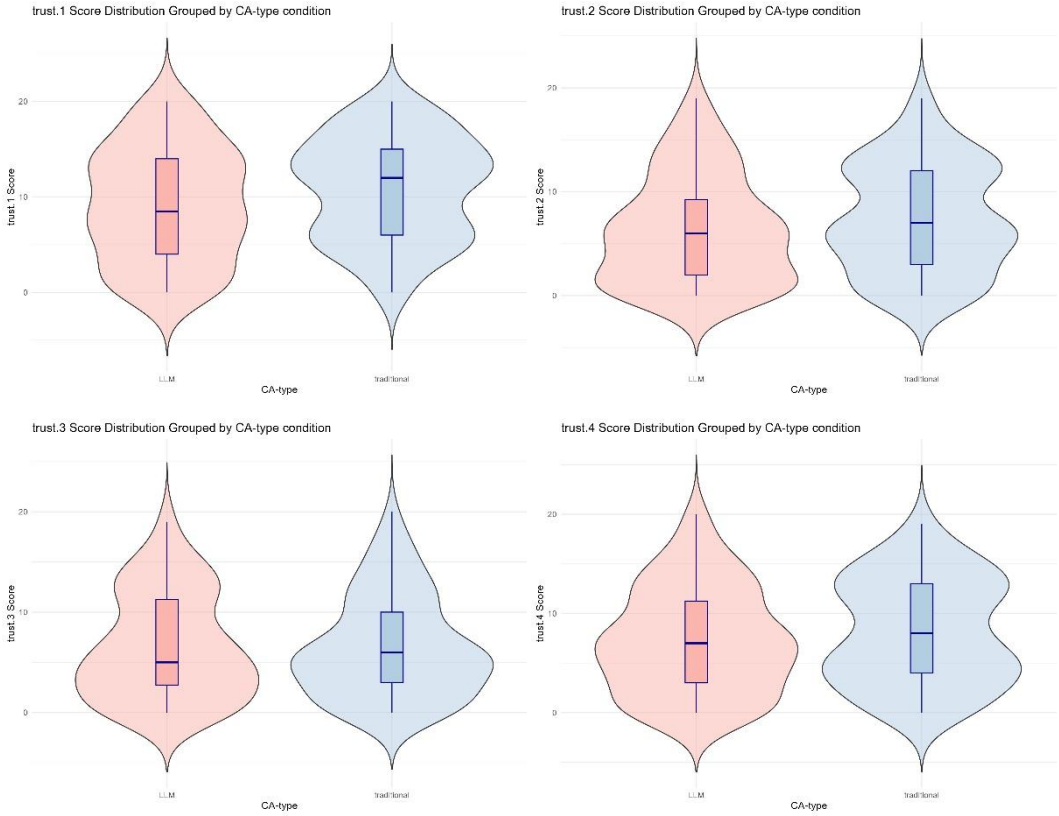


Figure.5 The comparison of four trust indicators' scores between two CA-type conditions.

This experiment further aggregated the mean values of the four trust indicators to represent an overarching trust score, which is called the average trust score. A Wilcoxon rank sum test has been conducted to assess the influence of CA-type conditions on this consolidated trust measure, but the difference is not statistically significant ($W = 2607.5$, $p = .075$).

In response to observed differences in trust.1 between interaction sequence groups, we conducted separated group analyses using the Wilcoxon signed-rank test, focusing on the impact of CA-type conditions on trust scores within each group. For Group A, the analysis showed a non-significant influence of CA-type on trust scores, as indicated by the statistics ($V = 317$, $p = .21$). Similarly, the analysis for Group B showed no significant effect of CA-type on trust scores ($V = 353.5$, $p = .32$).

H3: Relationship between autonomy and satisfaction

The same method used in this section; 6 Wilcoxon rank sum tests were used to calculate the score difference between CA-type conditions among 6 subscales of UEQ. Significant differences were observed in Attractiveness ($W = 2271.5$, $p < .005$), Perspicuity ($W = 2383.5$, $p = .01$), Efficiency ($W = 1962.5$, $p < .005$), Dependability ($W = 2336.5$, $p < .005$), and Stimulation ($W = 2480$, $p = .026$). Conversely, Novelty ($W = 3360.5$, $p = .40$) showed no significant difference by CA-type. However, Figure.6 illustrates that the TRA CA condition groups had a higher mean score than LLM CA condition groups in all the significant subscales.

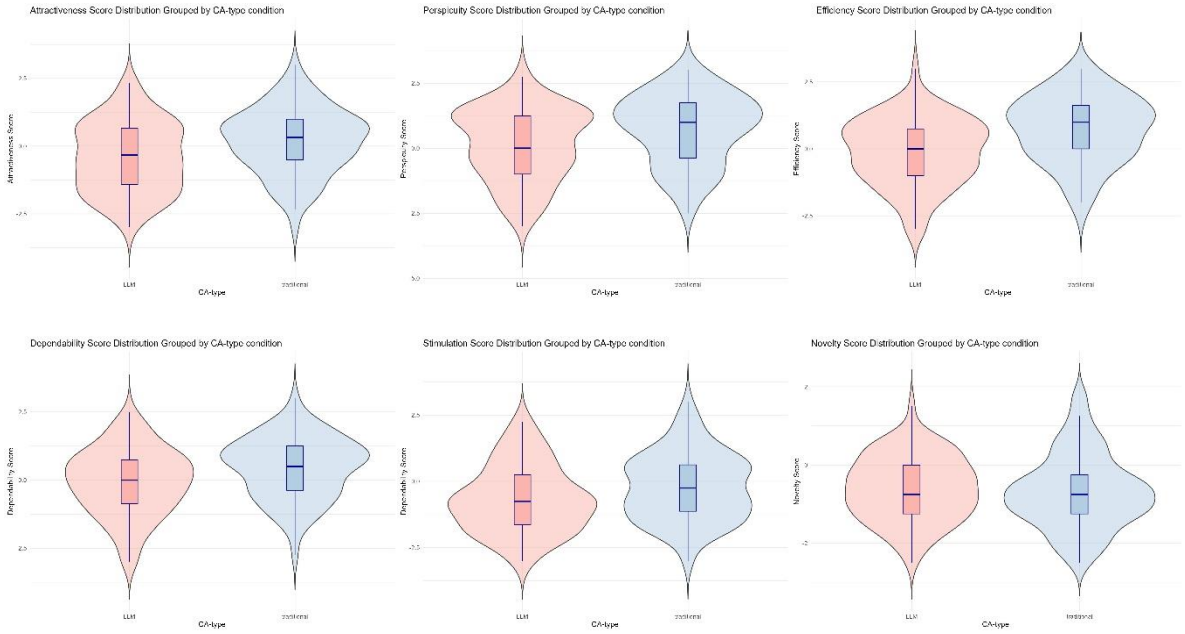


Figure.6 The comparison of six UEQ subscale scores between two CA-type conditions.

H4: Interaction effect of autonomy in the relationship between trust and satisfaction

In this section, the study used the average trust score to represent the participants' perceived "trust" and the Attractiveness as the representative of a general "satisfaction". This study used a linear regression model with interaction to investigate the relationship between the trust and attractiveness, and the interaction effect of CA-type condition in this relationship. The regression model demonstrated a strong fit with the data, $F (3, 154) = 101.6$, $p < .001$. Trust scores ($b = 0.22$, $SE = 0.018$, $t = 12.734$, $p < .001$) and CA-type (TRA) ($b = 0.50$, $SE = 0.24$, $t = 2.055$, $p = .042$) were significant predictors of Attractiveness. The interaction between trust scores and TRA CA type was not significant ($b = -0.019$, $SE = 0.026$, $t = -0.72$, $p = .47$). Attractiveness scores increased by 0.22 points for every one-point increase in trust scores and

by 0.5 points for cases identified as traditional Type, compared to other Types. The non-significant interaction suggests the effect of trust on attractiveness does not differ significantly between the TRA CA condition and the LLM CA condition. The left part of Figure.7 illustrates the overall data frame which combined group A & B, and the right part illustrates the data frames that are separated by the interaction sequence. From the figure, we can find that group B had greater significance on differences in the slope of the relationship between attractiveness and trust score (Group A only: b = 0.31, SE = 0.38, t = 0.82, p = .41; Group B only: b = 0.51, SE = 0.28, t = 1.86, p = .067), both of them have no significant interaction effect (Group A only: p = .47; Group B only: p = .24).

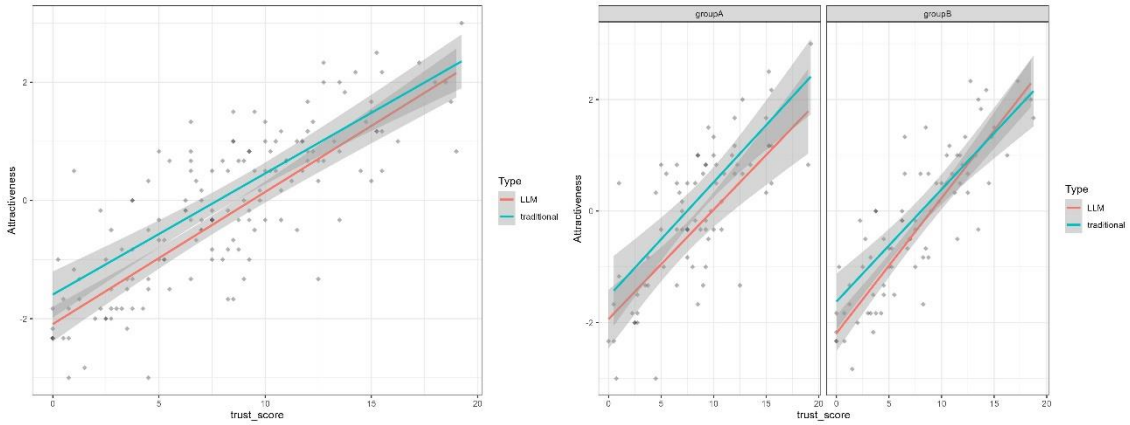


Figure.7 The relationship between Attractiveness and the average trust score grouped by CA-type conditions

4.2 Qualitative

4.2.1 Autonomy related - LLM CA condition

There are 107 meaningful reference codes in the open-ended questions about autonomy in LLM CA experiences. There were 58 references (54.21%) that showed negative attitudes; 8 references had negative attitudes toward the users themselves, and 36 references expressed negative attitudes toward the CA. Among these negative references, 15 of them related to the "incomprehension" and the "repetitiveness" of LLM CA:

"The only time this would be frustrating is if the CA does not understand your enquiry, and therefore kept trying to gather information that cannot help/you don't have access to."

Among 41 (38.32%) positive references, the majority of references (30 references) had a positive attitude toward the user themselves, and only 9 references related to the CA. 15 references of them related to the "free typing communication":

"It allows you to fully express your problems in detail. So you are able to control the amount of information it gives you based on the depth of the information that you give it."

4.2.2 Autonomy related - TRA CA condition

There were 102 meaningful reference codes in the open-ended questions about autonomy in TRA CA experiences. There were 66 references (64.71%) that showed negative attitudes; 8 references had negative attitudes toward the users themselves, and 55 references expressed negative attitudes toward the CA. Among these negative references, "limitation" is an essential theme. 34 references related to the dissatisfaction of interaction method, and 12 references complained that the user cannot choose to talk about the certain point that the user wants to ask. Meanwhile, there were another 12 references pointed to similar arguments of the option and direction provided by the TRA CA dialogue tree has a limited number, and the fixed path caused negative attitudes:

"As it was not a live CA and was pre-programmed, there were only certain permutations of paths that I could take. As such, I could not get to the specific question that I wanted to address as it is somewhat fixed and does not allow for more nuanced or unique inquiries."

Among 27 (38.32%) positive references, unlike the LLM CA condition, 17 references had positive attitude toward the CA, and only 7 references related to the user themselves. 6 references of them highlighted the efficiency, and 5 references pointed that the options acting as guidance in interaction:

"I felt like I had a bit of control over the discussion as I was guided by the different sections."

5. Discussion

5.1 Integration of the findings

Regarding Hypothesis 1, which investigated the influence of interaction sequence, the expectation was that none of the indicators would show significant differences between Group A and B. This hypothesis was only partially confirmed. Among the various indicators examined, three demonstrated significant differences: Trust.1, NASA.TLX.4, and Efficiency. These findings suggest that contrary to initial expectations, certain aspects of the interaction sequence do indeed have a significant impact on the user experience, particularly in terms of trust, perceived task load (as measured by NASA.TLX.4), and efficiency. The results highlight the importance of considering the sequence of interactions when designing and deploying such systems.

Hypothesis 2 explored the relationship between autonomy and the CA-type condition, with the expectation that Autonomy.1 would be significantly different from Autonomy.2 would show no difference. This hypothesis was fully supported by the findings. Autonomy.1 emerged as a robust indicator of "autonomy," validating the experimental design's effectiveness in controlling for the variable of "influence" and distinguishing it from "autonomy." This result is particularly significant as it demonstrates the experiment's success in isolating and examining the specific contribution of autonomy to the user experience in interacting with CAs.

Hypothesis 3 postulated that the LLM CA would be associated with higher trust and satisfaction scores compared to TRA CA conditions. Contrary to these predictions, the results indicated an unexpected trend in trust-related indicators: TRA conditions consistently outperformed LLM conditions across most indicators. Specifically, only one trust-related indicator (trust.1) showed a minor significance favouring LLM CA, while the remainder revealed no significant differences, often displaying an inverse relationship to the hypothesised direction. This pattern extended to the relationship between autonomy and

satisfaction, where, remarkably, 5 out of 6 indicators demonstrated significant differences between LLM and TRA conditions, yet again, contrary to the predicted superiority of LLM CAs.

Hypothesis 4 examined the interaction effect of autonomy on the relationship between trust and satisfaction, anticipating a significant positive correlation between these variables, with autonomy acting as a significant moderator. While trust was indeed positively correlated with satisfaction, affirming part of the hypothesis, the expected interaction effect of autonomy was not observed. Both conditions showed similar slopes in their relationships between trust and satisfaction, with the TRA CA condition exhibiting a higher intercept but no significant differences in the interaction effect of autonomy.

5.2 Interpretation of the discrepancy

Hypothesis 1: sequence effect

The observed differences in Group AB could be attributed to order effects. Combining the qualitative responses, this suggests that the sequence of interaction tasks influenced participant perception, at least in trust perception. This finding aligned with the concept of sequence effect mentioned by Cockburn et al. (2017). This acknowledgement of the limitations in controlling for order effects suggests a need for refined experimental designs that more effectively isolate the impact of the interaction task itself from the sequence in which it is presented.

Hypothesis 3: variable control, workload, familiarity

The primary concern identified in testing Hypothesis 3 relates to variable control, particularly regarding the workload assigned under different conditions involving LLM CA and TRA CA. The lack of strict control over the workload introduces variability that could confound the results, the study by Hillesheim et al. (2017) used the NASA TLX indicated that the high workload might be negatively correlated to the trust perception of users. Additionally, the experiment did not account for participants' familiarity with LLM interactions, nor did it provide tutorials for participants with low familiarity, potentially skewing results based on prior experience rather

than the intrinsic qualities of the conversational agents. Furthermore, design flaws and technical issues within the LLM CA condition may have adversely impacted participants' perceptions of trust and satisfaction (Salem et al., 2015; Desai et al., 2012). These oversights in experimental design and execution could have significant implications for the validity and reliability of the findings, suggesting that future studies should incorporate more rigorous control measures and consider participant familiarity and technical stability as critical factors in evaluating conversational agents.

Hypothesis 3: TRA CA guidance

According to the qualitative feedback, there was another explanation about how the user interacted with the CA and the role of TRA CA took in the conversation that is not considered in the previous experiment design and literature review. Some participants mentioned that the options provided by the TRA CA offered the user a guidance and inspiration when the user feel lost and have no direction about the next step in the interaction. Among the qualitative feedback, this opinion is not statistically general mentioned (4.9% references toward the TRA CA included this theme), but we still cannot estimate to what degree that participants' preference of CA was influenced by this condition since we have no relevant measurement that directly investigated on this direction.

Hypothesis 3: problem-solving & capability

The qualitative analysis revealed a stronger correlation between satisfaction and the capability of CA. There were 33.64% reference toward LLM CA expressed the negative attitude about the incompetence in the solving problems, but only 18.63% toward TRA CA. Meanwhile, 6.5% reference toward LLM CA recognised their experience of higher human autonomy, but that is not helpful in problem-solving. This outcome suggests that the capacity to effect change and solve problems, is a critical determinant of user satisfaction with an agent. This finding aligns with the notion that users value an agent's problem-solving capabilities highly, viewing them as a primary measure of the agent's utility and effectiveness (Følstad & Skjuve, 2019). The capacity to influence outcomes and navigate challenges efficiently appears to be more impactful on satisfaction than the mere freedom from external control, represented by

autonomy.

Hypothesis 3: technical & designing issues

Moreover, this study highlighted a limitation concerning the allocation of sufficient interaction time for participants. Specifically, the lack of adequate time disproportionately affected those with lower proficiency or less clarity in their approach, leading to incomplete experiments and, consequently, a negative experience for some. This aspect emphasises the importance of accommodating varying user skill levels and ensuring ample time for interaction to assess the agent's impact on user satisfaction fully. The correlation between influence and satisfaction, alongside the operational challenges encountered, provides valuable insights for future design and research efforts, particularly in optimising the balance between enabling user influence and ensuring comprehensive, accessible user experiences.

Another technical issue is about the misleading prompting design in the LLM CA. In order to replicate real-world corporate CA interaction patterns, we designed a mandatory "detail input condition", which required the LLM CA to gather certain information from the user. However, that caused a repetitive loop if the user was not familiar with the natural language processing interaction. Overall there were 17.76% of references toward LLM CA complained about this issue in the qualitative feedback. In these instances, participants' efforts to modify their inputs failed to elicit varied responses from the CA, trapping them in repetitive cycles. Despite LLMs' potential to afford users the liberty to input text freely, this persistent misalignment between user input and system response significantly undermined perceived autonomy. Effective autonomy in interaction design is not solely about offering the choice of action but also ensuring that these choices are meaningful and lead to discernible outcomes. Chan et al. (1995) have pointed out that system interaction without meaningful feedback will significantly deteriorate the user's performance and confidence.

Furthermore, it is imperative to recognise that while the challenges introduced in our simulation are meaningful, they do not fully capture the complexity of natural language interactions. The constraints applied to mimic certain interaction scenarios may fall short of

accurately representing the diverse and dynamic nature of real-world communication. This observation underscores the limitations inherent in our experimental setup and suggests that future research endeavours should aim for a higher degree of realism in modelling natural language interactions, thus enhancing the ecological validity of such studies.

Hypothesis 4: latent variable measurement

In our examination of the relationships between trust, satisfaction, and autonomy through linear regression, we encountered the complexity of accurately representing subjective perceptions with concrete indicators. Trust and satisfaction, as measured in our study, are latent variables derived from participants' subjective evaluations, raising concerns about the fidelity and reliability of the indicators used to encapsulate these constructs (Maslowsky et al., 2015). The questionnaire items related to trust were adapted from pre-existing research, which may have compromised their representativeness for our specific context. This adaptation process, while necessary for contextual relevance, introduces a potential limitation in accurately capturing the essence of trust as perceived in our study. Furthermore, our findings suggest that the relationships among trust, satisfaction, and their contributing factors may have been oversimplified. The observed limitations in our analysis of H3 carry significant implications for H4, given the hierarchical relationship between these two hypotheses. The challenges encountered in accurately modelling the dynamics of trust and satisfaction in H3 likely compromised the validity of subsequent analyses in H4.

5.3 Implication & future research direction

Future studies should prioritise assessing participants' familiarity with LLM and their pre-experiment attitudes towards such technologies, rather than focusing solely on their specific areas of expertise. This approach will allow for more nuanced control of variables present before the experiment, potentially transforming this assessment into a standalone study. It could explore how proficiency and predisposed preferences towards traditional applications (TRA CA in the content of this experiment) influence users' receptivity to, and the perceived

utility of, emerging technologies. This shift in measurement strategy aims to understand better the barriers to adoption and the negative impacts that unfamiliarity with new technologies might have on user experience.

Improving the representation of simulated challenges and the inclusion of realistic background information is crucial for better simulating real-life environments in experiments. Adjustments to these aspects could enhance the ecological validity of future studies, making the experimental conditions more reflective of the users' natural interactions with conversational agents.

Our findings hint at a potential mediation effect where autonomy influences satisfaction through the mediation of trust (Kassim et al., 2012). Future research should delve deeper into this structure, employing Confirmatory Factor Analysis to refine the measurement of these latent variables within our questionnaires. Further, applying Structural Equation Modelling would enable a more comprehensive analysis, considering the intricate relationships between trust, satisfaction, and autonomy as latent variables. This experiment reaffirmed the positive relationship between trust and satisfaction, also suggesting that autonomy might play a differential role in this relationship. The possibility of a mediation effect, where autonomy indirectly influences satisfaction through trust, warrants further investigation. Future experiments could employ structural equation modelling (SEM) to explore this mediation effect with enhanced variable control. Additionally, confirmatory factor analysis (CFA) could be utilised to validate the relationships among the constructs in our questionnaires, aiming for optimal measurement of the latent variables involved.

Through these future research directions, we aim to address the current study's limitations and pave the way for a deeper understanding of the complex interactions between users and conversational agents, ultimately contributing to the development of more user-friendly and effective HCI technologies.

6. Conclusion

The goal of this study was to investigate the influence of human autonomy on the user's perception of trust and satisfaction in the interaction with the conversational agent. This experiment used within-subject comparison indicated the participant who interacted with the LMM-powered conversational agent had a relatively higher status of human autonomy than the participant who interacted with the traditional branching dialogue conversational agent. Meanwhile, we measured the user's perception of trust and satisfaction when interacting with the different types of conversational agents, and the result indicated that the interaction with higher human autonomy status cannot be positively correlated with higher user perception in both trust and satisfaction. The findings also confirmed that among the user's perceptions, trust is positively correlated with satisfaction. Interestingly, human autonomy had no significant interaction effect on the relationship between trust and satisfaction.

According to the qualitative feedback and literature research, we find that the incongruent result might be caused by the order effect of the interaction sequence, the insufficient variable control on the workload of interaction tasks and user's familiarity levels, and the not considering the guidance effect on the dialogue options provided by the TRA CA. Meanwhile, this study found that the user had a higher weight on the CA's problem-solving capabilities rather than the autonomy. Moreover, the technical issues on interaction time and repetitive loops caused by the LLM CA prompting would also have a significant influence on the user's perception of trust and satisfaction towards different CAs. Lastly, the study argues that the relationship structure among the three concepts and the construct validity might need to be further considered. We suggest that future studies use the CFA to implement better indicators of concepts of user's perception of trust and satisfaction and use the SEM to test the mediation effect among autonomy, trust and satisfaction.

Reference

Anderson, J., Rainie, L., & Luchsinger, A. (2018). Artificial intelligence and the future of humans. *Pew Research Center*, *10*(12).

Apraiz, A., Lasa, G., & Mazmela, M. (2023). Evaluation of User Experience in Human–Robot Interaction: A Systematic Literature Review. *International Journal of Social Robotics*, *15*(2), 187-210.

Ariely, D. (2000). Controlling the information flow: Effects on consumers' decision making and preferences. *Journal of consumer research*, *27*(2), 233-248.

Bailly, G., Lecolinet, E., & Nigay, L. (2016). Visual menu techniques. *ACM Computing Surveys (CSUR)*, *49*(4), 1-41.

Balaji, D. (2019). Assessing user satisfaction with information chatbots: a preliminary investigation (Master's thesis, University of Twente).

Benbasat I, Wang W. Trust in and adoption of online recommendation agents. Journal of the association for information systems. 2005;6(3):4.

Bentham, J. (1789). An introduction to the principles of morals. Athlone.

Bokhari, R. H. (2005). The relationship between system usage and user satisfaction: a meta-analysis. Journal of Enterprise Information Management, 18(2), 211-234.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Calvo, R. A., Peters, D., Vold, K., & Ryan, R. M. (2020). Supporting human autonomy in AI systems: A framework for ethical enquiry. *Ethics of digital well-being: A multidisciplinary approach*, 31-54.

Chan, H. C., Wei, K. K., & Siau, K. L. (1995). The effect of a database feedback system on user performance. *Behaviour & Information Technology*, *14*(3), 152-162.

Cichocki, A., & Kuleshov, A. P. (2021). Future trends for human-ai collaboration: A comprehensive taxonomy of AI/AGI Using Multiple Intelligences and Learning Styles. *Computational Intelligence and Neuroscience*, *2021*, 1-21.

Cockburn, A., Quinn, P., & Gutwin, C. (2017). The effects of interaction sequencing on user experience and preference. *International Journal of Human-Computer Studies*, *108*, 89-104.

Deci, E. L., & Ryan, R. M. (2000). The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, *11*(4), 227-268.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., ... & Yanco, H. (2012,

March). Effects of changing reliability on trust of robot systems. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 73-80).

Ellison, B. (8). July 2008." Defining Dialogue Systems". Gamasutra.

EU, A. (2021). European Commission white paper on artificial intelligence–a European approach to excellence and trust. *URL https://ec. europa. eu/info/sites/info/files/commissionwhitepaper-artificial-intelligence-feb2020 en. pdf*.

Følstad, A., & Skjuve, M. (2019, August). Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st international conference on conversational user interfaces* (pp. 1-9).

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In International Conference on Internet Science, 194-208. Springer, Cham.

Fraser, J., Papaioannou, I., & Lemon, O. (2018, November). Spoken conversational ai in video games: Emotional dialogue management increases user engagement. In *Proceedings of the 18th international conference on intelligent virtual agents* (pp. 179-184).

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.

Grand View Research. (2021). Chatbot Market Size, Share & Trends Analysis Report By End User, By Application (Bots For Service, Bots For Marketing), By Type (Standalone, Web-based), By Product Landscape, By Vertical, And Segment Forecasts, 2021–2028.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis*, *11*(3), 255-274.

Gupta, K., Hajika, R., Pai, Y. S., Duenser, A., Lochner, M., & Billinghurst, M. (2020, March). Measuring human trust in a virtual assistant using physiological sensing in virtual reality. In *2020 IEEE Conference on virtual reality and 3D user interfaces (VR)* (pp. 756-765). IEEE.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology*, *25*(2), 91-97.

Hillesheim, A. J., Rusnock, C. F., Bindewald, J. M., & Miller, M. E. (2017, September). Relationships between user demographics and user trust in an autonomous agent. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 314-318). Sage CA: Los Angeles, CA: SAGE Publications.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of

large language models. *arXiv preprint arXiv:2307.10169*.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological science*, *4*(6), 401-405.

Kassim, E. S., Jailani, S. F. A. K., Hairuddin, H., & Zamzuri, N. H. (2012). Information system acceptance and user satisfaction: The mediating role of trust. *Procedia-Social and Behavioral Sciences*, *57*, 412-418.

Kopp, S. (2006). How people talk to a virtual human-conversations from a real-world application. *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, *101*.

Lammers, J., Stoker, J. I., Rink, F., & Galinsky, A. D. (2016). To have control over or to be free from others? The desire for power reflects a need for autonomy. *Personality and Social Psychology Bulletin*, *42*(4), 498-512.

Lammers, J., Stoker, J. I., Rink, F., & Galinsky, A. D. (2016). To have control over or to be free from others? The desire for power reflects a need for autonomy. *Personality and Social Psychology Bulletin*, *42*(4), 498-512.

Lankton, N., McKnight, D. H., & Thatcher, J. B. (2014). Incorporating trust-in-technology into Expectation Disconfirmation Theory. Journal of Strategic Information Systems, 23, 128–145.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4 (pp. 63-76). Springer Berlin Heidelberg.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Lucia-Palacios, L., & Pérez-López, R. (2023). How can autonomy improve consumer experience when interacting with smart products?. *Journal of Research in Interactive Marketing*, *17*(1), 19-37.

Manaris, B. (1998). Natural language processing: A human-computer interaction perspective. In *Advances in Computers* (Vol. 47, pp. 1-66). Elsevier.

Maslowsky, J., Jager, J., & Hemken, D. (2015). Estimating and interpreting latent variable interactions: A tutorial for applying the latent moderated structural equations method. *International journal of behavioral development*, *39*(1), 87-96.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The journal of strategic information systems*, *11*(3-4), 297-323.

Mill, J. S. (1861). Representative government. Kessinger Publishing.

Mostafa, R. B., & Kasamani, T. (2022). Antecedents and consequences of chatbot initial trust. European journal of marketing, 56(6), 1748-1771.

Nasirian, F., Ahmadian, M., & Lee, O. K. D. (2017). AI-based voice assistant systems: Evaluating from the interaction and trust perspectives.

Ogden, W. C., & Bernick, P. (1997). Using natural language interfaces. In *Handbook of human-computer interaction* (pp. 137-161). North-Holland.

Paap, K. R., & Cooke, N. J. (1997). Design of menus. In *Handbook of human-computer interaction* (pp. 533-572). North-Holland.

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019). Better language models and their implications. *OpenAI blog*, *1*(2).

Rauschenberger, M., Olschner, S., Cota, M. P., Schrepp, M., & Thomaschewski, J. (2012, June). Measurement of user experience: A Spanish language version of the user experience questionnaire (UEQ). In 7th Iberian Conference on Information Systems and Technologies (CISTI 2012) (pp. 1-6). IEEE.

Roca, J. C., & Gagné, M. (2008). Understanding e-learning continuance intention in the workplace: A self-determination theory perspective. *Computers in human behavior*, *24*(4), 1585-1604.

Rödel, C., Stadler, S., Meschtscherjakov, A., & Tscheligi, M. (2014, September). Towards autonomous cars: The effect of autonomy levels on acceptance and user experience. In *Proceedings of the 6th international conference on automotive user interfaces and interactive vehicular applications* (pp. 1-8).

Ross, S. I., Martinez, F., Houde, S., Muller, M., & Weisz, J. D. (2023, March). The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 491-514).

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of management review*, *23*(3), 393-404.

Ryan, R. M.,&Deci, E. L.(2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford publications.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015, March). Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 141-148).

Sankaran, S., Zhang, C., Gutierrez Lopez, M., & Väänänen, K. (2020, October). Respecting human autonomy through human-centered AI. In Proceedings of the 11th Nordic Conference on Human-Computer Interaction:

Shaping Experiences, Shaping Society (pp. 1-3).

Schoonenboom, J., & Johnson, R. B. (2017). How to construct a mixed methods research design. Kolner Zeitschrift fur Soziologie und Sozialpsychologie, 69(Suppl 2), 107.

Shteingart, H., Neiman, T., & Loewenstein, Y. (2013). The role of first impression in operant learning. *Journal of Experimental Psychology: General*, *142*(2), 476.

Smestad, T. L. (2018). *Personality Matters! Improving The User Experience of Chatbot Interfaces-Personality provides a stable pattern to guide the design and behaviour of conversational agents* (Master's thesis, NTNU).

Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter for information systems users. *European Journal of Information Systems*, *25*(3), 274-287.

Taylor-Giles, L.C. (2014). *Toward a deeper understanding of branching dialogue systems* (Doctoral dissertation, Queensland University of Technology).

Tyack, A., & Mekler, E. D. (2020, April). Self-determination theory in HCI games research: Current uses and open questions. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-22).

Ulfert, A. S., Antoni, C. H., & Ellwart, T. (2022). The role of agent autonomy in using decision support systems at work. *Computers in Human Behavior*, *126*, 106987.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. Nature Human Behaviour, 7(9), 1526-1541.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Yen, C., & Chiang, M. C. (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), 1177-1194.

Zemčík, M. T. (2019). A brief history of chatbots. *DEStech Transactions on Computer Science and Engineering*, *10*.

Appendix:

Appendix.1: The background information and simulated challange

The background scenario

Scenario Overview: As a participant in this experiment, you will step into the shoes of a new tenant who has just moved into a property at the end of October 2023. Your task is to navigate a realistic situation involving utility suppliers by interacting with a conversational agent (CA) from an energy company. Here is your background story:

Your New Home and the Utility Mystery: You recently moved into your new home, excited about starting a fresh chapter. In the utility room of your new property, you discovered a note left behind by the previous tenant. This note contained an account number and password for a company named Insight Energy, along with a message stating, "Top up this account; it should cover your utility usage." Trusting this information, you didn't give it a second thought.

The Hot Water Dilemma: One day, you encountered an issue: the hot water supply suddenly stopped. Curious and concerned, you visited Insight Energy's website and realized that the account credit had run out. After topping up the account, the hot water supply was thankfully restored.

The Unexpected Bill: Life went on smoothly until you received an unexpected energy bill from another company, LGC. This jogged your memory – before moving in, you had researched utility suppliers in the region and learned that LGC was the primary provider. Acting on this information, you had registered an account on LGC's website but had since forgotten about it, due to the note from the previous tenant indicating Insight Energy as your supplier.

Confusion and Concern: Now, you find yourself in a perplexing situation. Two energy companies appear to be charging for utilities at your property. With bills from both Insight Energy and LGC, you are understandably confused and concerned about whom you should be paying.

Seeking Support: To resolve this confusion and get some clarity, you decide to seek support. You turn to the LGC website, hoping to find answers and guidance through their conversational agent.

Next

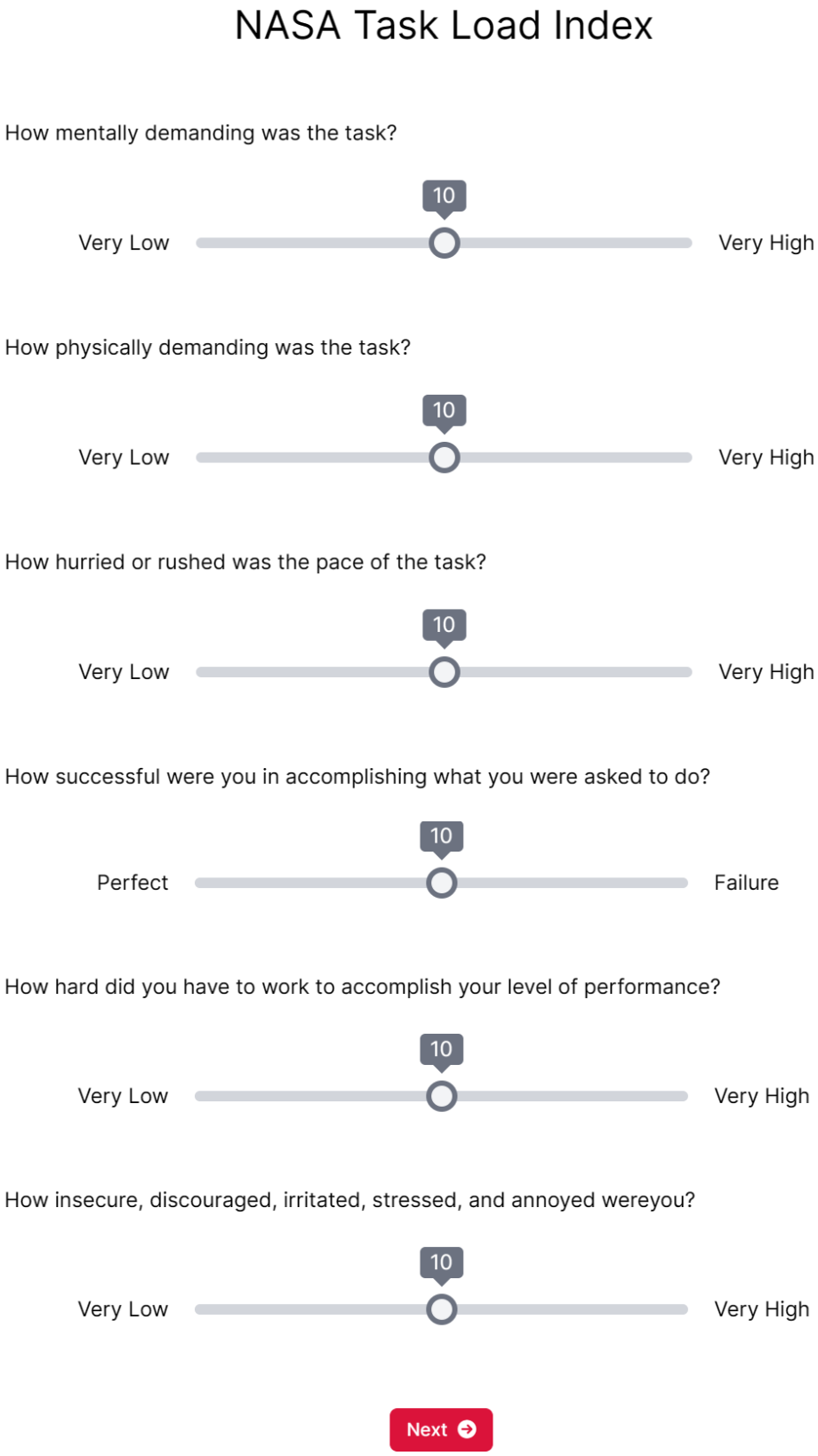
LGC Energy Bill

- Billing Covering: 28 October 2023 - 28 November 2023
- Bill Date: 28 November 2023
- Account Number: A12345
- Property Postcode: A1 2BC
- Total Energy Cost (Excluding VAT): £60
- VAT at 5%: £3
- Your New Balance on 28 November 2023: £63

Next

Appendix.2: Interaction phase questionnaires

The NASA TLX



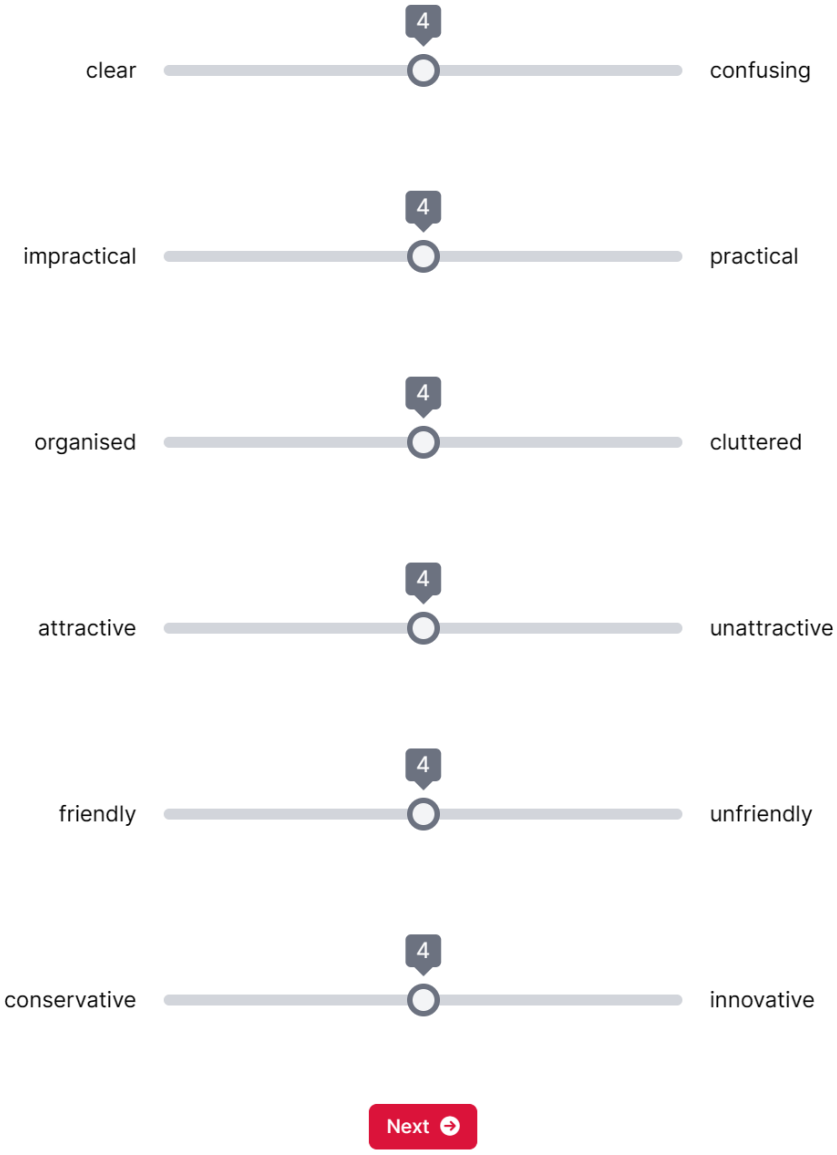
The UEQ

User Experience Questionnaire

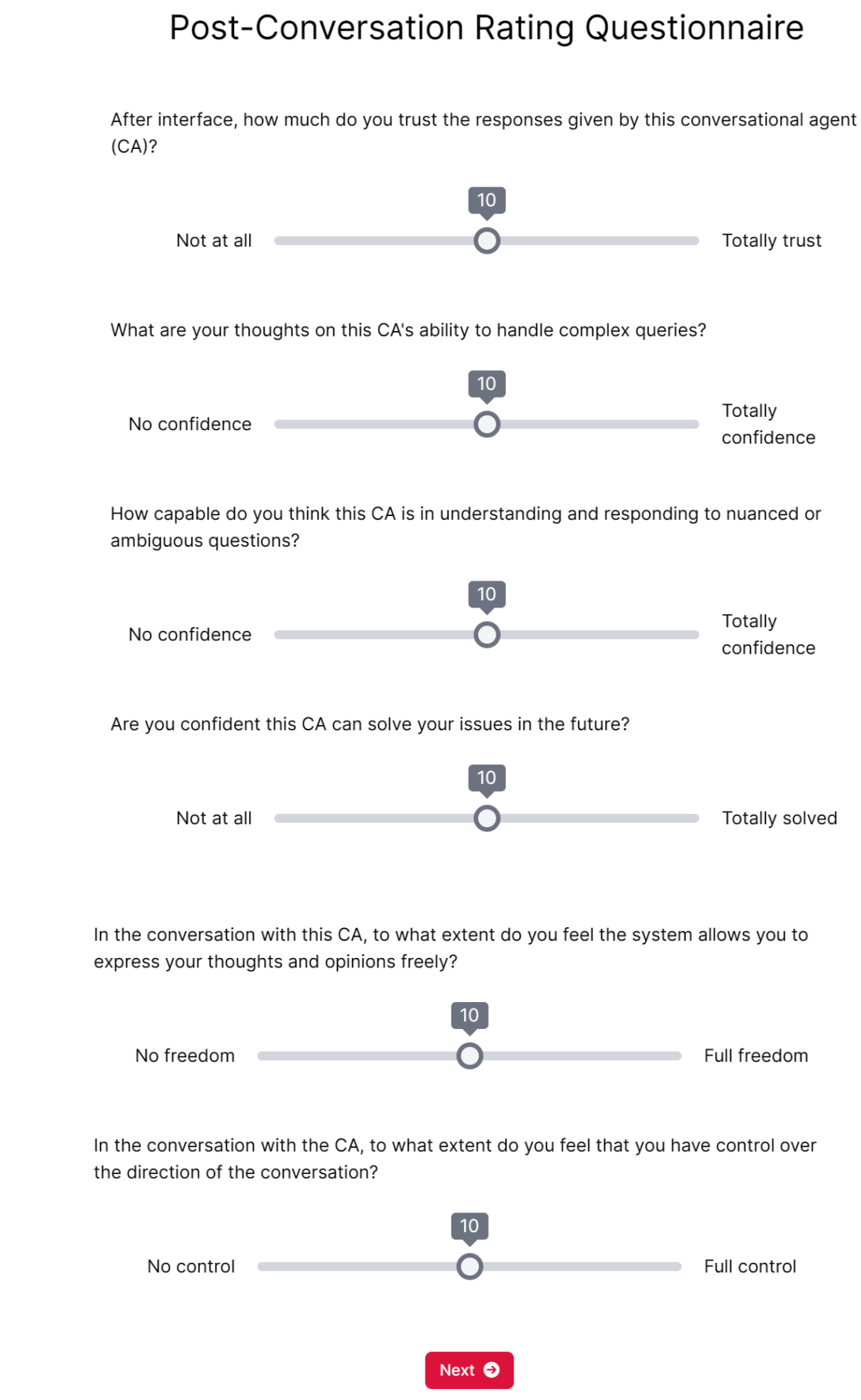
For the assessing the experience of the "Conversational Agent", please fill out the following questionnaire. The questionnaire consists of pairs of contrasting attributes that may apply to the product. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by ticking the circle that most closely reflects your impression.

Please assess the product now by ticking one circle per line.


inventive
4
conventional
obstructive
4
supportive
good
4
bad
complicated
4
easy
unlikable
4
pleasing
usual
4
leading edge
unpleasant
4
pleasant
secure
4
not secure
motivating
4
demotivating
meets expectations
4
does not meet expectations
inefficient
4
efficient


4
clear
confusing
4
impractical
practical
4
organised
cluttered
4
attractive
unattractive
4
friendly
unfriendly
4
conservative
innovative
Next

The self-rating question collection



The open-ended question collection

Open-ended Question (Feelings)

The questions are open-ended, designed to elicit detailed responses about your experiences and perceptions. There are no right or wrong answers, and we encourage you to be as honest and thorough as possible.

Feel free to elaborate on your answers as much as you like. The more detail you provide, the better we can understand your experiences. However, there is no strict word account limitation.

1. "How do you feel about your ability to control the direction of information gathering when interacting with the conversational agent? Please explain."

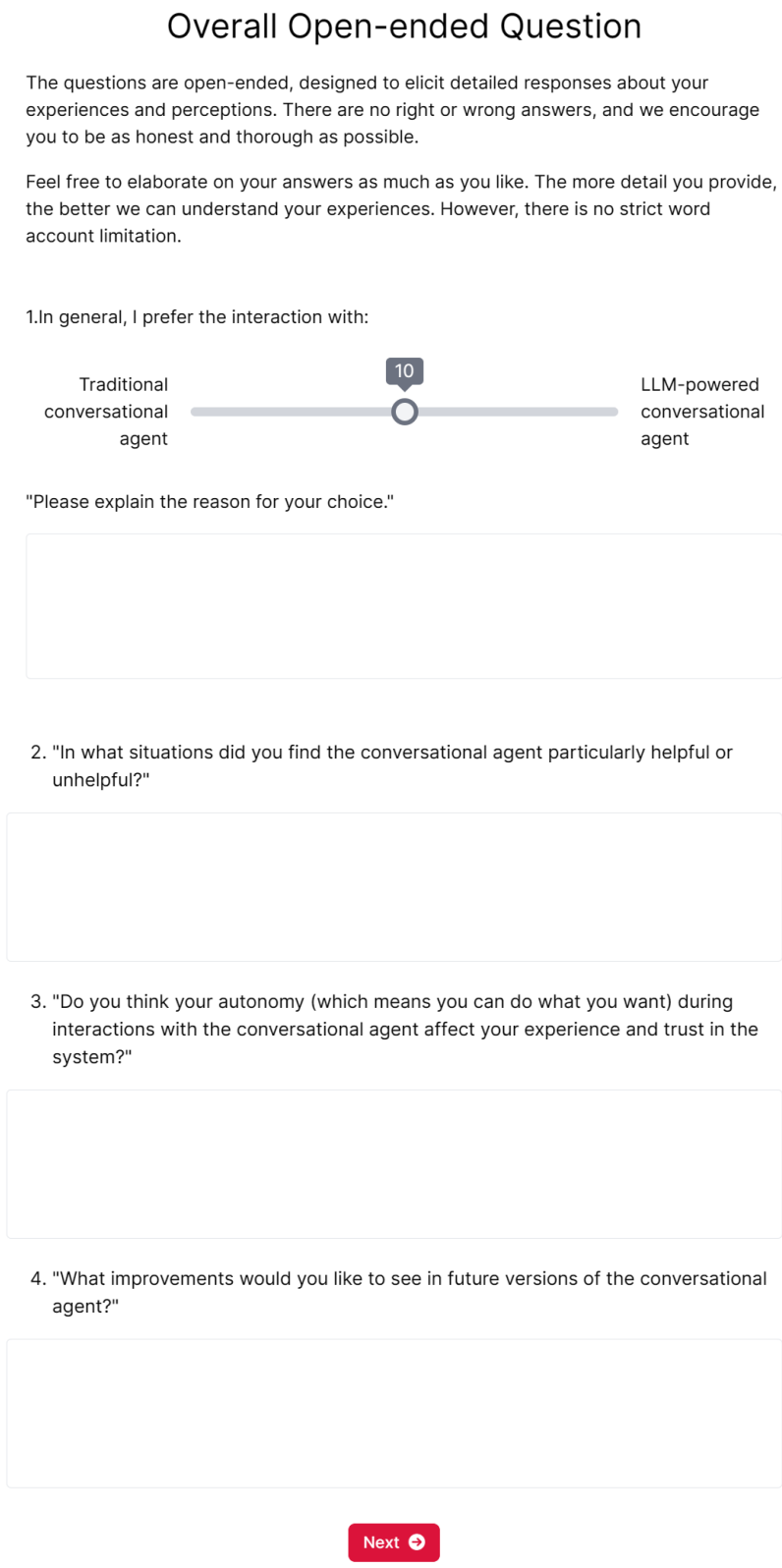
2. "In what situations did you find the conversational agent particularly helpful or unhelpful?"

3. "Describe any emotional responses you have had while interacting with the conversational agent."

Next

Appendix.3: Preference & demographic phase questionnaires

The preference



The demographic questionnaire

Demographic & Experiences

What is your age group?

- ◯ Under 18
- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ 55 and above

What is your gender?

- ◯ Male
- ◯ Female
- ◯ Non-binary/Third gender
- ◯ Prefer not to say
- ◯ Prefer to self-describe:

What is the highest level of education you have completed?

- ◯ High school and below
- ◯ Undergraduate
- ◯ Postgraduate
- ◯ PhD and higher

What is your current occupation or field of work? (Please select the option that best describes your area. If you are a student, please select the field of your major study.)

- ◯ Technology (e.g., IT, Software Development, Engineering)
- ◯ Healthcare (e.g., Medicine, Nursing, Healthcare Administration)
- ◯ Education (e.g., Teacher, Professor, Educational Administration)
- ◯ Business (e.g., Management, Finance, Marketing)
- ◯ Creative Arts (e.g., Art, Design, Entertainment)
- ◯ Science and Research (e.g., Biology, Physics, Social Sciences)
- ◯ Government or Public Service
- ◯ Trades and Services (e.g., Construction, Plumbing, Electrical)
- ◯ Retail or Customer Service
- ◯ Hospitality (e.g., Food Service, Hotel Management)
- ◯ Other - please specify:

Next