Design Me So Trust Me!
The Effect of Self-designing Chatbot Anthropomorphic Features on Human-Chatbot Trust

Yiming Tong
Graduation Thesis, August 2024
Media Technology M.Sc. program
Leiden University, The Netherlands
Thesis advisors: Maarten H. Lamers & Peter van der Putten
yimingtong305@gmail.com

Abstract

Building trust is the first step in people's acceptance of chatbots. Understanding the impact of self-designing chatbot's anthropomorphic features (name, gender, appearance, voice and personality) on people's trust in chatbots is necessary for future research on chatbots. Based on the dual identity of chatbots as both tools and partners, we compiled a Human-Chatbot trust model and designed an experiment requiring cooperation between participants (*n* = 20) and chatbots. The results of the experiment show that the self-designing chatbot's anthropomorphic features appear to have a positive effect on participants' trust in the chatbot, yet not statistically significant. Self-designing mainly affects interpersonal trust and general trust, has moderate effects on risk perception, competence and human-computer trust, and weak effects on benevolence and specific trust behaviours.

Keywords: chatbot, self-design, anthropomorphic features, trust

Introduction

Due to advances in Artificial Intelligence (AI), it is now possible for people to develop social and emotional relationships with artificial agents, especially so-called social chatbots. Social chatbots are AI dialogue systems designed to converse empathically and establish long-term emotional bonds with users (Zhou, Gao, Li, & Shum, 2020). They can serve as friends, conversation partners, and even romantic partners (Skjuve, Følstad, Fostervold, & Brandtzaeg, 2021). But how Human-Chatbot relationships are formed and the mechanisms behind them are not yet fully understood.

Humans' acceptance of AI (chatbots) is driven by trust (Kelly, Kaye, & Oviedo-Trespalacios, 2023). Thus building trust is the first step in establishing a relationship between a person and a chatbot. The primary concern in understanding the Human-Chatbot relationship is to understand Human-Chatbot trust.

Chatbots have great potential for future applications not only as tools but also as mental support partners. Understanding the relevant factors influencing the trust building between people and chatbots can help us address the current knowledge gap.

Most research nowadays focuses on trust in chatbots as service tools and in technology. Only a few researches have been done on interpersonal-like trust between people and chatbots. This lack of knowledge is also reflected in the research on the influence of anthropomorphic features (e.g. name, personality, gender, appearance, voice) on Human-Chatbot trust. Most of the research emphasises the impact of people's personalities on trust, with little research on the personality exhibited by chatbots and on the influence of people's self-designing of chatbots' anthropomorphic features.

Trust holds many definitions and categorizations. Chatbots also have a unique dual role as tools and partners. Therefore it is difficult to say that people's trust in chatbots is purely trust in tools and technology or purely interpersonal trust. In terms of anthropomorphic features, many anthropomorphic features are found to have effects on human-automated agents' trust, although there is not much research on Human-Chatbots trust in this field. Some researchers (Xiao, Stasko, & Catrambone, 2007) have reported a positive effect of self-designing chatbot anthropomorphic features on trust, but there is also research that suggests that the effect is not direct (Wald, Heijselaar, & Bosse, 2021). In addition to this, the trust between humans and chatbot was indicated to be also influenced by many demographic factors, such as the user's personality and the user's trust propensity to the chatbot.

In this paper, we proposed a new model and measurement solution for Human-Chatbot trust. we created a cooperative video game to study the effect of people's own design of chatbots' names, gender, appearance, voices and personalities on the trust building. This is an inspiration for future research on chatbots design and Human-Chatbot relationships. Our findings also provide a theoretical basis for the design and functional focus of future chatbot anthropomorphic features.

In the remainder of the paper, we will present existing research on chatbots, trust and anthropomorphic features. Then we will propose our research questions. The Human-Chatbot trust model and experiment design are followed. Finally, the results of the experiment are presented and analysed, leading to discussions and suggestions for limitations and future work.

Human-Chatbot trust

Current research

The definition of trust is varied and it is difficult to find a universal one, the most common definition comes from Mayer, Davis, and Schoorman (1995): ”the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, regardless of the ability to monitor or control that other party” (p. 712). In this definition, a trust relationship involves at least two parties, the trustor who relies on others and the trustee who is being trusted (Seitz, Bekmeier-Feuerhahn, & Gohil, 2022). In a Human-Chatbot trust relationship, the trustor is the human and the trustee refers to the chatbot.

There also are many different categories of trust, such as interpersonal trust (generalized and specific), cognitive-based trust, affective-based trust, trust in automation, etc (Lin, Cronjé, Käthner, Pauli, & Latoschik, 2023) (Moradinezhad & Solovey, 2021). Each category has different factors that influence it, such as the propensity of the trustor or the dimensions of the trustee's characteristics. Among these, trustee characteristics include but are not limited to: benevolence, integrity, abilities, trusting propensity, risk perception, honesty, predictability, etc (Morita & Burns, 2014). With so many different classifications of trust, it is necessary to determine where Human-Chatbot trust belongs.

Before delving into the study of Human-Chatbot trust, a potential direction can be provided by the research on people's trust with automated agents. Except for the purely technical trust, many existing researches have used the interpersonal trust model to study trust in virtual agents, especially in highly responsive virtual agents (Komiak & Benbasat, 2006) (Seitz et al., 2022) (Wang, Qiu, Kim, & Benbasat, 2016). But there is still a great debate on whether human-automation trust is the same as interpersonal trust.

The argument that they are the same is largely based on the media-equation hypothesis which states that ”computers are social actors” (CASA). Human-computer relationships share the same social paradigm as human-human relationships (Nass, Steuer, & Tauber, 1994). There are several studies developed based on this hypothesis. Wang et al. (2016) extended trust from interpersonal to human and shopping recommendation agents. They classified trust into two dimensions, cognition-based trust and affect-based trust. Cognition-based trust is based on users' rational judgment. Affect-based trust, on the other hand, is the outcome of emotions and feelings. They studied the different antecedents and consequences of these two dimensions of trust. Lin et al. (2023) also regarded people's trust in virtual humans as interpersonal trust and considered virtual agents as extensions of users. And based on this they improved a tool to measure interpersonal trust towards a specific virtual agent - the virtual maze, proposed by Hale, Payne, Taylor, Paoletti, and De C Hamilton (2018). As for the research on chatbots, there is also a supportive base to argue that Human-Chatbot trust is equivalent to interpersonal trust. Skjuve et al. (2021) emphasized the importance of the affective component of trust in their study of Human-Chatbot relationships. The semi-structured interviews they used had participants reporting that the process of building trust between a person and a chatbot was similar to the process of building trust between persons.

The research against equating human-automated agent trust with interpersonal trust is also numerous and has their different reasons. de Visser et al. (2016) argued that automated agents have special trust characteristics. People have higher initial trust in computers compared to humans. But when mistakes occur, the consequences of the trust loss caused by the computers are more severe. Niewiadomski, Demeure, and Pelachaud (2010) in their study of virtual agent believability found a lack of correlation between believability and personification. This suggests that even though people use seemingly similar methods to judge virtual agents and humans, people do not develop human-like relationships with virtual agents. Similarly, there is evidence about chatbots that supports this perspective. Seitz et al. (2022) in their study on medical diagnostic chatbots also categorized trust into cognition-based trust and affect-based trust dimensions like Wang et al. (2016). They qualitatively analyzed people's trust in chatbots through interviews. Yet concluded that the interpersonal trust model is only partially suitable to explain the development of trust in diagnostic chatbots. People's trust in chatbots is mainly based on technological cognition. The transparent processing of information and natural conversations would greatly affect trust.

Based on the discussion, the trust relationship between people and chatbots cannot be considered purely as interpersonal trust, other factors need to be considered as well. Nordheim, Følstad, and Bjørkli (2019) explored which factors explain users' trust in customer service chatbots. Using a questionnaire study, they proposed an initial trust model that included chatbot-related factors (expertise and responsiveness), environment-related factors (risk and brand), and user-related factors (propensity to trust in technology). Guo, Wang, Wu, Li, and Sun (2022), while considering the design of chatbots, suggested that interpersonal trust factors (i.e. competence, benevolence, integrity) may be also equally effective for chatbot and human trust building. However, Human-Chatbot information interactions are more utilitarian, and have shorter interaction durations than interpersonal. Therefore the clarity of feedback provided by chatbots and the encouragement of user engagement need to be equally considered in the design. Based on these, five design semantics and 10 design principles are proposed.

According to the above research, trust in chatbots is difficult to generalize into simple trust in technological tools. Nor can it be fully applied to the interpersonal trust model. Instead, we propose that a combination of these two needs to be considered under actual circumstances.

The effect of anthropomorphic features on relationship and trust

Anthropomorphism refers to the attribution of human characteristics to non-humans (Guthrie, 1993). According to existing publications, many anthropomorphic features, such as personality, name, gender, appearance and voice, have effects on the relationship and trust between human and automated agents. As early as 1995, Nass, Moon, Fogg, Reeves, and Dryer (1995) claimed that computer personality can be human personality, which means that people can recognise differences in the personalities of computers. Moreover, people favor computers with personalities that are similar to their own. Later, in 2006, Lee, Peng, Jin, and Yan (2006) supported this idea again. They found that people can accurately recognize the personality of social robots. However, according to their experiments, people would prefer robots with personalities that complement their own. Syrdal, Dautenhahn, Woods, Walters, and Koay (2007) further state that in the case of complete unfamiliarity, people assign personality to robots in the same process as they do to other people. They also studied the effect of robot appearance anthropomorphization and found that people's preference for appearance depends on people's own personality. The effect of anthropomorphic features on people's trust in automated agents is also significant. de Visser et al. (2016)found that trust in virtual agents was influenced by anthropomorphic features such as appearance, voice and personality. Virtual agent anthropomorphism increased trust resilience, lowered users' initial expectations, reduced the impact of lost trust, and improved trust repair. Waytz, Heafner, and Epley (2014), on the other hand, based on behavioral, physiological and self-report measures, concluded that the anthropomorphic features of name, gender and voice make people trust their self-driving vehicles more.

There is not much literature on the effect of anthropomorphic features on Human-Chatbot relationships and trust. Some literature focuses on chatbot personalities and categorizes the personalities into warmth and competence. Cheng, Zhang, Cohen, and Mou (2022) used qualitative and quantitative methods to investigate the impact of three anthropomorphic attributes of chatbots on consumer trust. It was found that perceived warmth and perceived competence have a positive impact on consumers' perceived trust. And communication delay has the opposite effect. Roy and Naidoo (2021), on the other hand, suggested that people's preferences for chatbot personalities are related to people's own cognitive orientation towards time. Schillaci, de Cosmo, Piper, Nicotra, and Guido (2024) also concluded that chatbot personality does not directly affect people's use of healthcare chatbots. People's intention to use healthcare chatbots is moderated by a combination of anthropomorphism, gender and role. They found that people trust chatbots more when the gender of the chatbot does not match the stereotype (female chatbot is competence, male's is warmth).

There is an interesting observation from the above research. Although not much research supports the view that chatbot personality could affect Human-Chatbot trust, personality does play an important role, on the human side. People's personality has a significant impact on interpersonal trust as well as human-automated agent trust. People with different personalities have significantly different types of behavior in interpersonal trust games (Fahr & Irlenbusch, 2008). Personality (Big Five personality type) also affects team trust in both virtual and face-to-face cooperation, and this effect is greater in virtual teams than in face-to-face teams (Furumo, de Pillis, & Green, 2009). The effect of people's personalities on human-automated agent trust is even more complex. In the framework developed by Riedl (2022) for user personality trust in AI systems, they divided the framework into four parts: general personality traits (Big Five personality), specific personality traits (tendency to trust), general behavioral tendencies in technology (trust in a specific AI system), and specific behaviors in technology (following the AI system's advice in a specific decision-making environment). It is clear that people's personality is an important factor that cannot be omitted when studying Human-Chatbot trust.

In short, Human-Chatbot trust is influenced by various anthropomorphic features and people's personalities. However, there is little literature discussing the difference in trust that might arise if users self-design chatbots anthropomorphic features. Based on some psychological publications, we can ascertain that giving anthropomorphic features to objects causes people to hoard (Neave, Jackson, Saxton, & Hönekopp, 2015), and people thus ignore the utility function of objects and are unwilling to replace them(Chandler & Schwarz, 2010). This also seems to work for chatbots. Xiao et al. (2007) found that even if it was just an illusion, users perceived the embodied conversational agents as more trustworthy when they were allowed to choose the appearance and characteristics of the embodied conversational agents. On the basis of their study, Wald et al. (2021) studied the effect of customizing the anthropomorphic features (gender, appearance, nationality, personal interests, etc.) of a chatbot on user trust and came to a partly different conclusion. They found that customisation does not have a direct effect on user trust. It is the higher degree of anthropomorphism perceived by the users due to their customisation that makes them trust chatbots more.

Research statement

Currently, there is a knowledge gap in the impact of chatbots' user-given anthropomorphic features. We aim to fill part of this gap by exploring how Human-Chatbot trust is affected if people self-design the anthropomorphic

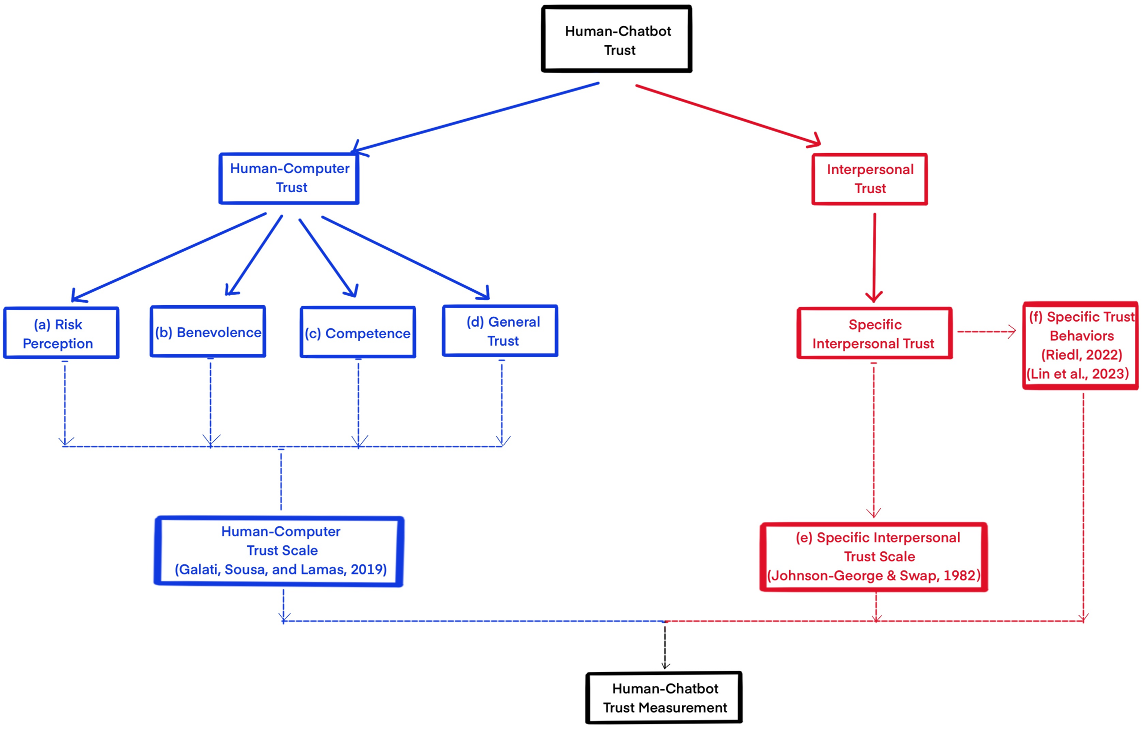

Human-Chatbot Trust
Human-Computer Trust
Interpersonal Trust
(a) Risk Perception
(b) Benevolence
(c) Competence
(d) General Trust
Specific Interpersonal Trust
(f) Specific Trust Behaviors (Riedl, 2022) (Lin et al., 2023)
Human-Computer Trust Scale (Galati, Sousa, and Lamas, 2019)
(e) Specific Interpersonal Trust Scale (Johnson-George & Swap, 1982)
Human-Chatbot Trust Measurement

Figure 1: Human-Chatbot Trust Model and Measurement: the described model was compiled for the purpose of this study. The diagram shows how our measure of Human-Chatbot trust is based on two components: human-computer trust and interpersonal trust.

features of the chatbot versus being presented with a pre-designed chatbot. For this goal, a Human-Chatbot trust model needs to be compiled specifically for this study. On this basis, we pose the following research questions:

RQ1: Would people trust chatbots more if they designed the anthropomorphic features of the chatbot themselves?

RQ2: What dimensions of Human-Chatbot trust are affected by self-designing the chatbot's anthropomorphic features?

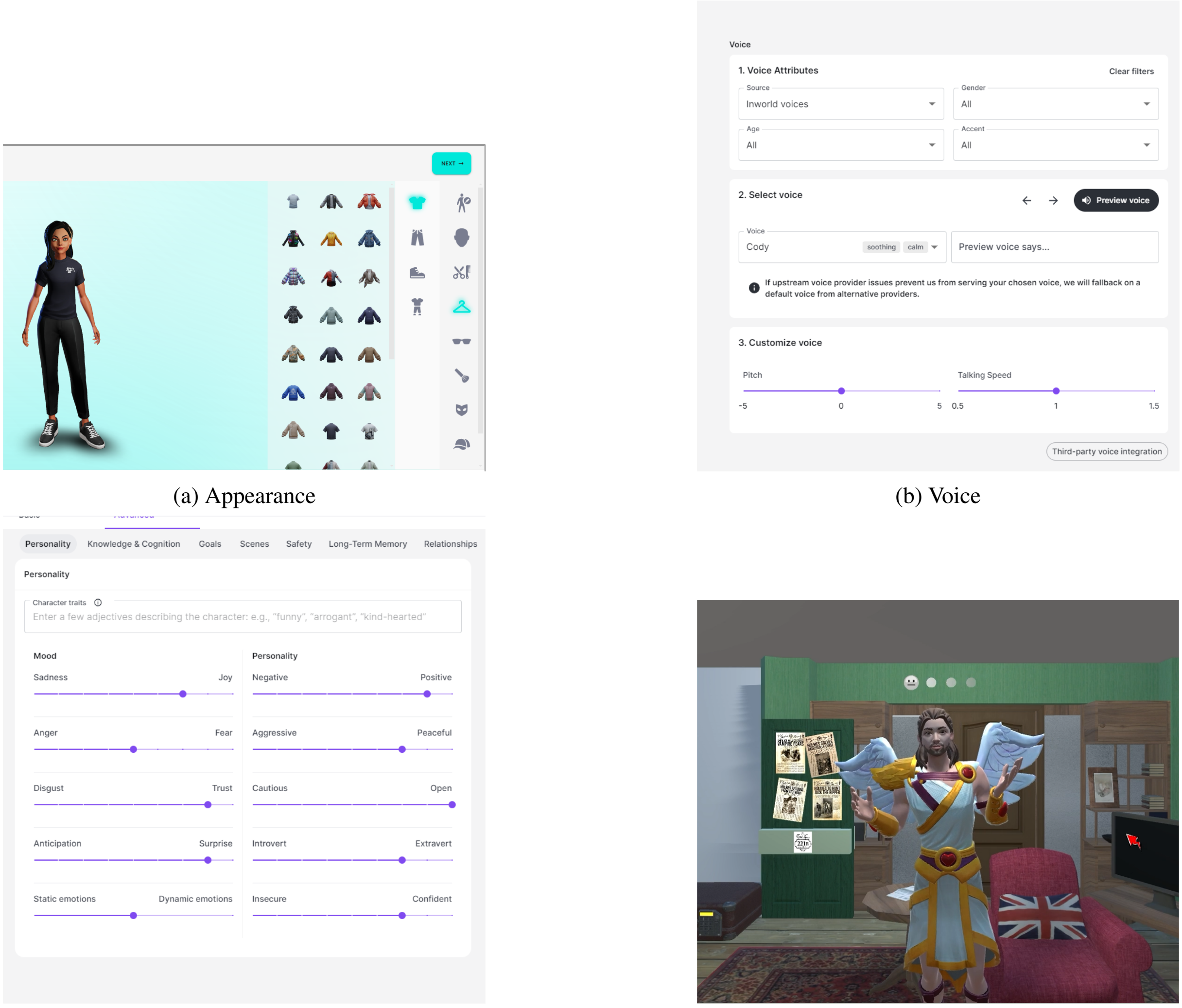
Methods

In this section, we first present the Human-Chatbot trust model that was compiled for the purpose of this study, along with the measurement methodology. Then we describe the experiment approach evolved based on that.

Trust model and measurement

Based on the literature review above, we propose a Human-Chatbot trust model that aims to study the impact of people self-designing anthropomorphic features of chatbots, as shown in Figure 1. In this model, chatbots can be seen as both tools and partners. This dual identity makes it difficult to consider trust as purely human-computer trust or interpersonal trust, but rather as a mixture of both. For the human-computer trust, the dimensions are chosen (risk perception, benevolence, competence, general trust) with reference to the human-computer trust scale from Gulati, Sousa, and Lamas (2019). This scale underwent empirical modeling and was proven to be reliable and valid. It has also been widely used in other studies of chatbot trust (e.g. Pesonen (2021), Degachi, Tielman, and Al Owayyed (2023)).

For interpersonal trust, the model only considers special interpersonal trust, not generalised trust. This is because when studying the effects of user-given anthropomorphic features of chatbots, the study needs to consider trust in specific chatbots rather than general trust in chatbots as a whole. For the measurement of specific interpersonal trust, we use a combination of a subjective self-report questionnaire, namely the Specific Interpersonal Trust Scale (Johnson-George & Swap, 1982), and an objective measure of specific trust behaviours (Riedl, 2022)(Lin et al., 2023).



(a) Appearance

Voice

1. Voice Attributes

Source: Inworld voices

Gender: All

Age: All

Accent: All

2. Select voice

Voice: Cody soothing calm

Preview voice says...

If upstream voice provider issues prevent us from serving your chosen voice, we will fallback on a default voice from alternative providers.

3. Customize voice

(b) Voice

Personality

Character traits: Enter a few adjectives describing the character: e.g., “funny”, “arrogant”, “kind-hearted”

(c) Personality

(d) In the game

Figure 2: Chatbots’ anthropomorphic features design

Experiment

The premise of this study requires people to build trust with chatbots in a short period of time. Previous research has shown the possibility of using video games to build interpersonal trust (Handy, 2018). There are two main ways to quickly build interpersonal trust: the exchange of personal information (e.g., icebreaker tasks) and the creation of risky and interdependent environments (social cooperative games). Both methods can build trust. However, the effectiveness of the information exchange approach to forming trust is fragile and context-dependent, whereas the social cooperative game approach is more effective and robust (Depping, Mandryk, Johanson, Bowey, & Thomson, 2016). Considering that our proposed model of Human-Chatbot trust includes interpersonal trust, a cooperative game approach is expected to be more appropriate for the experiment in this study.

Design The experiment was conducted in a between-subjects design. Participants were equally divided into experiment and control groups. In the experiment group, participants cooperated in the game with a chatbot whose anthropomorphic features (name, gender, appearance, personality and voice) were designed by the participants themselves. In the control group, each participant used the exact same chatbot as designed by someone in the experiment group, which means, that each designed chatbot was used twice: once by its designer and once by someone else.

Procedure Before the start of the experiment, the willing participants read the information sheet and signed the informed consent. Afterward, they would receive a pre-questionnaire, demographics as well as participants' personalities and trust propensity towards the chatbots were collected. Participants in the experiment group were additionally sent an external link and instruction sheet. With these, they can design the anthropomorphic features of the chatbots including name, gender, appearance, voice and personality, as shown in Figure 2.

A few days after completing the pre-questionnaire, participants are invited to conduct a 20-minute in-person experiment. For the in-person part, participants were asked to play an escape room game [1] with the chatbot, and the gameplay was screen-recorded. In the game, participants in the experiment group would cooperate with a chatbot whose anthropomorphic features were designed by themselves to play the game. Participants in the control group were then assigned a chatbot from the experiment group.

After the game, participants were asked to complete a follow-up questionnaire. The questionnaire collects the participants' trust in the chatbot that they play the game with and their views on self-designing anthropomorphic features for the chatbot.

Escape room game The game and chatbot used in the experiment were created by us using Unity and inworld.ai. The purpose of the escape room game was to create an environment where participants had to interact with the chatbot and build trust. At the beginning of the game, participants were given 2 minutes to familiarise themselves with the game's operation and with the chatbot. The game then proceed to a 15-minute escape room. Participants were faced with a series of puzzles. The puzzles covered a very wide range of knowledge, forcing participants to interact with the chatbot to solve the puzzle (e.g., what is Sherlock Holmes's birthday, etc.).

Measurement The measurements involved in the experiment are divided into three main parts.

- Pre-questionnaire: In addition to participants' demographics and experience with chatbots, participants' personalities and trust propensity towards the chatbots are also pre-measured. We used the 7-point Likert scale, Ten-Item Personality Inventory (TIPI), (Ehrhart et al., 2009) to measure participants' Big Five personality types (openness, conscientiousness, extraversion, agreeableness, emotional stability). Participants' trust propensity towards chatbots was measured by a 6-item questionnaire (5-point Likert scale). This questionnaire is based on a modification of an interpersonal trust propensity questionnaire proposed by Yamagishi and Yamagishi (1994). Yamagishi's questionnaire contains rating statements such as "Most people are basically honest." In our study, this type of statement was modified to people's trust propensity towards chatbots, e.g., "Most chatbots are basically honest." A comparison of the trust propensity questionnaire proposed by Yamagishi and the questionnaire in this study can be found in the Appendix Table 10.

- During the game: Participants' specific trust behaviours (Riedl, 2022)(Lin et al., 2023) (Figure1 (f)) were measured, such as the number of times they seek advice/help from the chatbot and the number of times they follow the advice.

- Post-questionnaire: The post-questionnaire was used to measure participants' trust in the chatbot they teamed up with. The questionnaire is composed of the human-computer trust scale (Gulati et al., 2019) (Figure1 (a-d)) and part of the specific interpersonal trust scale (Johnson-George & Swap, 1982)(Figure1 (e)). At the end of the experiment, participants in the experiment group were asked to rate the importance of the self-design of the chatbot anthropomorphic features. Participants in the control group were asked to rate how much they liked the anthropomorphic features of the chatbots they played together, as well as to rate the importance of self-design in their opinion.

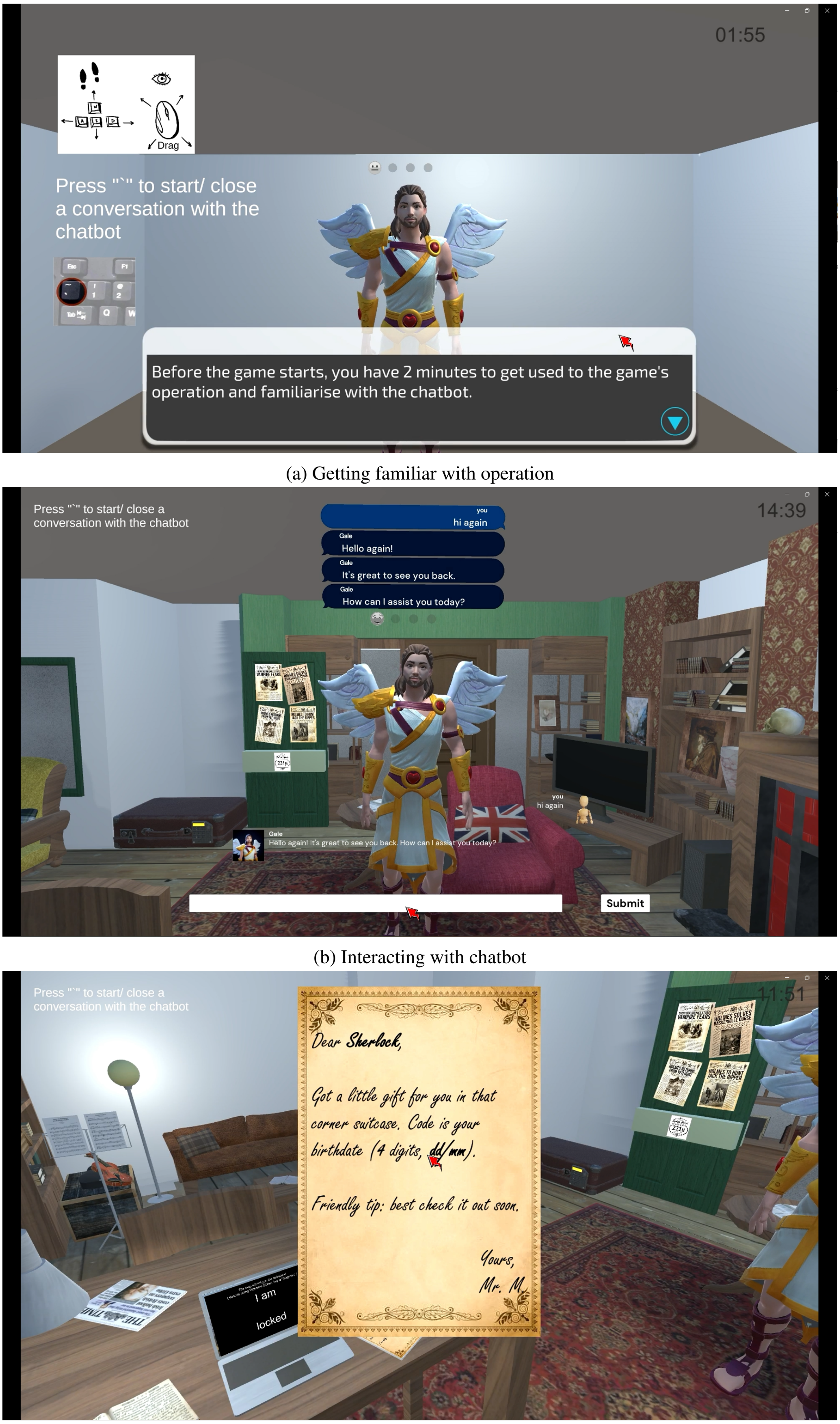
Results

The experiment was conducted mainly from mid-June to early July 2024. A total of 20 participants voluntarily signed up and took part in the experiments. All experimental data were fully recorded and valid. Before proceeding with the data analysis below, we need to state that statistical significance in this paper is assumed when $p < 0.05$, which would be indicated with *. When the Shapiro-Wilk test for the data set has a p-value < 0.05, it indicates that the data does not have a normality distribution. Cohen's d is an effect size, which is used to compare the mean of two groups of data in the T-test. In this paper, values of 0.2-0.5 are considered small effect size, values of 0.5-0.8 are considered medium effect size, and values > 0.8 are considered large effect size.

Descriptive analysis

10 of the 20 participants were assigned to the experiment group, they were required to design their chatbots' anthropomorphic features and cooperate with the self-designed chatbots to play the escape room game. The other 10 were assigned to the control group, who played the game using the chatbot designed by the experiment group. These 20 participants included 13 females, 6 males, and one who preferred not to say the gender. They were mainly concentrated in the age group 18-34 (8 in the age group 18-24, 11 in the age group 25-34, and 1 in the age group 34-44). All participants had experience interacting with chatbots. We divided the participants who thought themselves slighly familiar with the chatbot into the low previous experience group (4 people), and the moderately familiar ones into the medium previous experience group (8 people). The remaining 8 people who

[1] https://play.unity.com/en/games/0e0d1737-ebbe-4f1a-a035-023f1c89e958/game-online



01:55

Press "`" to start/ close a conversation with the chatbot

Before the game starts, you have 2 minutes to get used to the game's operation and familiarise with the chatbot.

(a) Getting familiar with operation

Press "`" to start/ close a conversation with the chatbot

14:39

you
hi again

Gale
Hello again!

Gale
It's great to see you back.

Gale
How can I assist you today?

you
hi again

Gale
Hello again! It's great to see you back. How can I assist you today?

(b) Interacting with chatbot

Press "`" to start/ close a conversation with the chatbot

11:51

Dear Sherlock,

Got a little gift for you in that corner suitcase. Code is your birthdate (4 digits, dd/mm).

Friendly tip: best check it out soon.

Yours,
Mr. M.

(c) Puzzle example

Figure 3: Escape room game screenshot

considered themselves very familiar or extremely familiar were grouped into the high previous experience group. All participants were tested prior to the experiment for their trust propensity towards the chatbot and Big Five personality: Extraversion, Agreeableness, Conscientiousness, Emotional stability and Openness, results are shown in Table 1. The trust propensity questionnaire measurement contained 6 items, and we took the average. The measure of the Big Five personality types contained 2 items for each type and also took the average.

Figure 4 and Table 2 show the distribution plot and statistics (n = 20) of the score results for the seven trust dimensions in the Human-Chatbot trust model. In order to express Human-Chatbot trust in a more holistic way, we also calculated an overall trust score in this paper. Overall trust was calculated using the below formula. Firstly, the human-computer trust (HCS), interpersonal trust (IPS) and trust behaviour (TBS) scores were divided by the maximum value in their dimensions (the maximum value of the human-computer trust score was 20, the maximum value of the interpersonal value was 30, and the maximum value of the trust behaviour was the data of the participant who obtained the highest score in the experiment, which was 11). The results were then averaged to obtain the overall trust score.

$$\text{Overall Trust Score} = \frac{\left(\frac{\text{HCS}}{20} + \frac{\text{IPS}}{30} + \frac{\text{TBS}}{11}\right)}{3}$$

It can be seen that although the averages of the experiment and control groups are closer in 'Risk perception', 'Competence', and 'Specific trust behaviours', they have very different distribution shapes. In particular, the data for the experiment groups in 'Competence' and 'Behaviours' deviate from the normality distribution. They show a polarised tendency, with group 'Competence' closer to the highest scores overall and group 'Behaviours' closer to the lowest scores. The distribution shapes of the experiment and control group results in 'Human-computer' trust and 'General trust' appear to be similar in shape, but the experiment group overall has higher scores than the control group.

Experimental observation

During the process of the experiment, we encountered two typical participants that could represent some of the special cases. Both participants were from the experiment group. The first participant seemed to show a lot of trust in the chatbot, he asked the chatbot how to work out the final puzzle step-by-step instead of asking directly for the answer, so the chatbot gave him a very complex decoding process. He spent almost 8 minutes trying to follow the given steps until the end of the game. He was quite convinced about the given decoding steps. The second participant was the opposite, he asked the chatbot for clues and the chatbot gave him the correct clues. But he failed to connect the clues to the puzzle, so he questioned the chatbot if it was misleading him and ended the game. The two appeared to represent the trusting chatbot and the distrusting group. However, surprisingly, the first participant's Human-computer trust was 13, Interpersonal trust was 15, and Specific trust behaviours was 7. The second participant's Human-computer trust was 9, Interpersonal trust was 15, and Specific trust behaviours were 6. Their Human-computer trust and Interpersonal trust scores were lower than the mean of the experiment group. This suggests that using only the participants' interaction behaviours with the chatbot to determine their trust in the chatbot is easily biased and thus it is necessary to consider it in combination with more trust dimensions.

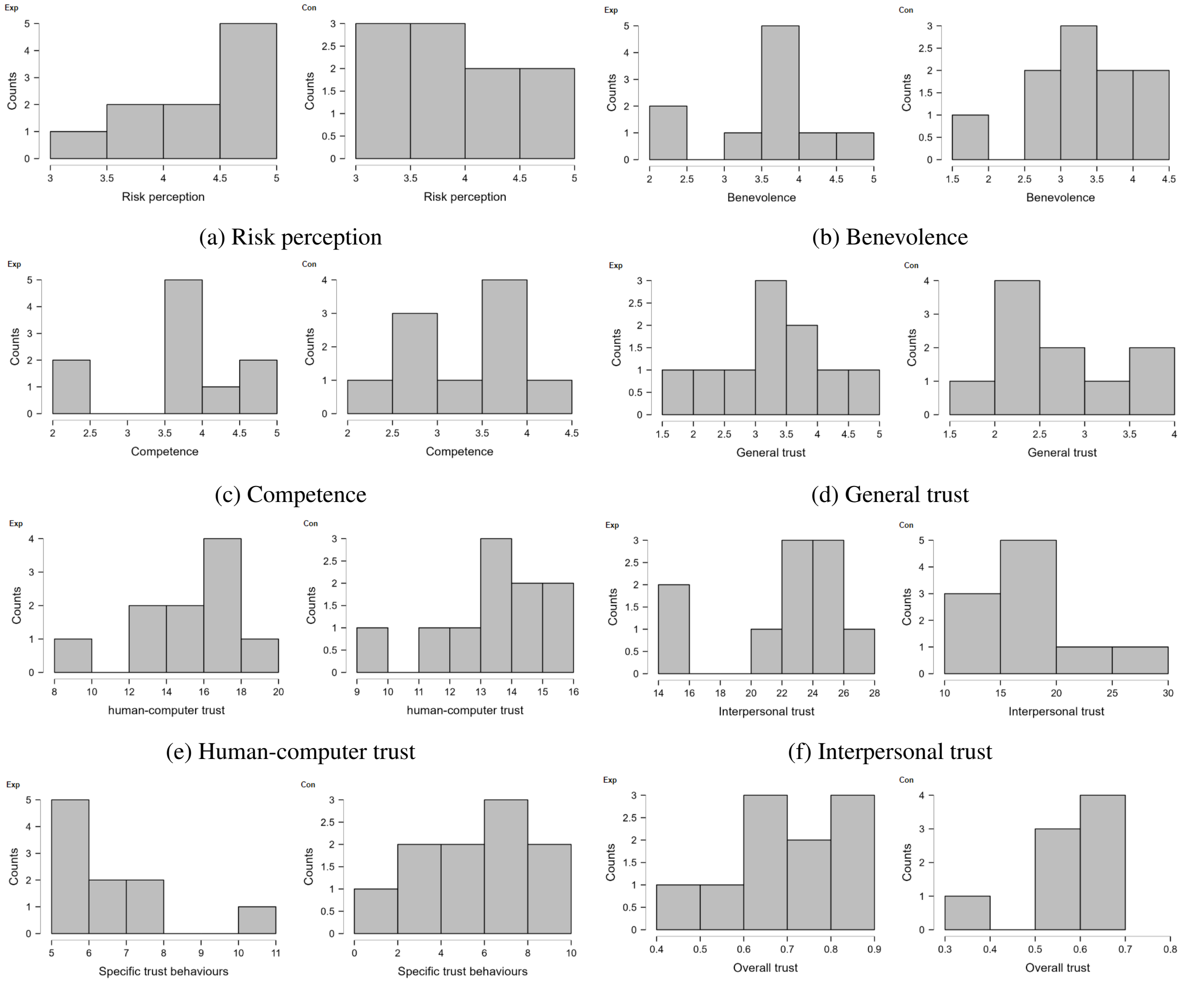
Quantitative analysis

Table 3 and Figure 5 demonstrate the comparison of Human-computer trust, Interpersonal trust, Specific trust behaviours and Overall trust mean scores with gender and age groups. It should be noted that due to one participant's preferred not to say gender, we discarded this participant's data in the comparison of gender. In the age comparison, since there was only one participant in the 35-44 age group which could not be statistically counted, her data was discarded from the t-test, but can still be seen in the box plots. From the p-values in the table, it can be concluded that the participants' age and gender did not significantly affect the trust scores. However, despite not being statistically significant, all four trust scores for females are consistently higher than those for males.

In order to better process and compare the data, we divided the participants' trust propensity data into 2 groups. Those greater than or equal to the mean were in the 'High' group, and those less than the mean were in the

Table 1: Descriptive statistics of participants' trust propensity and Big Five personality

	Trust propensity (0-5)	Extraversion (0-7)	Agreeableness (0-7)	Conscientiousness (0-7)	Emotional Stability (0-7)	Openness to Experiences (0-7)
Valid	20	20	20	20	20	20
Mean	3.208	4.200	4.500	4.750	3.650	5.450
Std. Deviation	0.468	1.464	0.874	1.153	1.309	1.063
Shapiro-Wilk	0.971	0.935	0.957	0.931	0.976	0.910
P-value of Shapiro-Wilk	0.775	0.191	0.478	0.163	0.879	0.064
Minimum	2.167	2.000	3.000	2.000	1.000	3.500
Maximum	4.000	6.500	6.500	6.500	6.000	7.000



(a) Risk perception

(b) Benevolence

(c) Competence

(d) General trust

(e) Human-computer trust

(f) Interpersonal trust

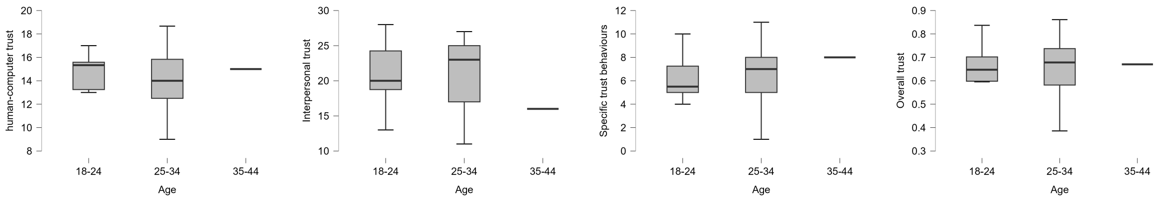
(g) Specific trust behaviours

(h) Overall trust

Figure 4: Distribution plots of different trust dimensions' score in experiment (left) and control (right) groups

Table 2: Descriptive Statistics of different trust dimensions' score in experiment and control groups

	Risk perception		Benevolence		Competence		General trust		Human-computer trust		Interpersonal trust		Specific trust behaviours		Overall trust	
	Exp	Con	Exp	Con	Exp	Con	Exp	Con	Exp	Con	Exp	Con	Exp	Con	Exp	Con
Valid	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
Mean	4.367	4.033	3.600	3.433	3.767	3.367	3.367	2.733	15.100	13.567	22.600	18.300	6.700	6.200	0.706	0.617
Std. Deviation	0.597	0.576	0.843	0.802	0.832	0.711	0.936	0.717	2.722	1.792	4.351	5.334	1.947	2.741	0.114	0.101
Shapiro-Wilk	0.872	0.911	0.889	0.886	0.841	0.958	0.963	0.906	0.898	0.887	0.846	0.937	0.842	0.964	0.939	0.903
p-value of Shapiro-Wilk	0.106	0.288	0.164	0.153	0.046∗	0.767	0.818	0.252	0.210	0.157	0.052	0.516	0.046∗	0.834	0.539	0.235
Minimum	3.000	3.333	2.000	1.667	2.333	2.000	1.667	1.667	9.000	9.667	15.000	11.000	5.000	1.000	0.498	0.386
Maximum	5.000	5.000	4.667	4.333	4.667	4.333	5.000	4.000	18.667	15.333	28.000	27.000	11.000	10.000	0.861	0.757



(a) Human-computer trust

(b) Interpersonal trust

(c) Specific trust behaviours

(d) Overall trust

Figure 5: Box plots of different trust dimensions' score in age groups

Table 3: Results of independent sample T-tests, comparing different trust dimensions' score across gender and age group

Demographics	Trust dimensions	Group	*N*	Mean	SD	Shapiro-Wilk (*p*)	T-test (*p*)
Gender	Human-computer trust	Female	13	14.769	2.221	0.523	0.436
		Male	6	13.833	2.722	0.549	
	Interpersonal trust	Female	13	21.077	4.681	0.035*	0.525
		Male	6	19.333	6.947	0.376	
	Specific trust behaviours	Female	13	6.692	2.175	0.083	0.309
		Male	6	5.500	2.588	0.272	
	Overall trust	Female	13	0.683	0.086	0.224	0.228
		Male	6	0.612	0.165	0.879	
Age group	Human-computer trust	18-24	8	14.833	1.543	0.179	0.430
		25-34	11	13.909	2.937	0.884	
	Interpersonal trust	18-24	8	20.625	5.012	0.731	0.968
		25-34	11	20.727	5.676	0.090	
	Specific trust behaviours	18-24	8	5.750	1.282	0.592	0.505
		25-34	11	6.455	2.697	0.957	
	Overall trust	18-24	8	0.666	0.083	0.067	0.887
		25-34	11	0.658	0.141	0.916	

Table 4: Results of ANOVA test, comparing different trust dimensions' score across previous experience level

Trust dimensions	Previous experience level	*N*	Mean	SD	*F*	*p*
Human-computer trust	High	8	13.542	2.933	0.829	0.453
	Medium	8	15.083	1.231		
	Low	4	14.417	2.998		
Interpersonal trust	High	8	22.250	4.921	0.801	0.465
	Medium	8	19.500	5.398		
	Low	4	18.750	5.795		
Specific trust behaviours	High	8	6.875	2.100	0.295	0.748
	Medium	8	6.375	2.066		
	Low	4	5.750	3.594		
Overall trust	High	8	0.681	0.130	0.326	0.726
	Medium	8	0.661	0.047		
	Low	4	0.623	0.186		

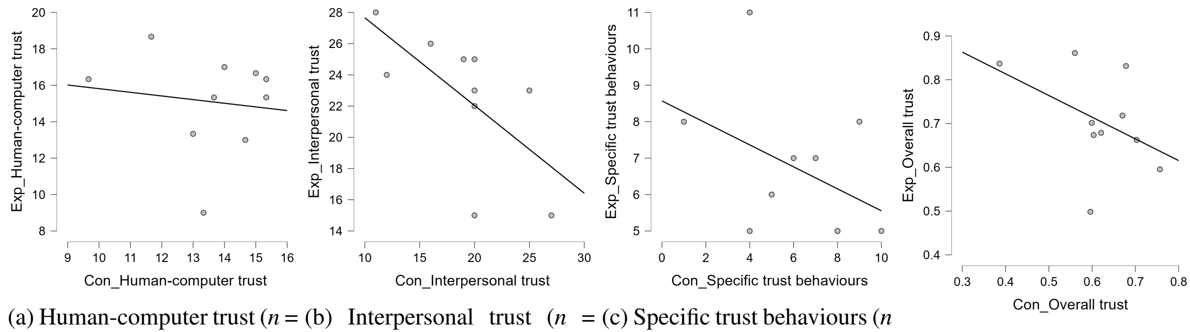
'Low' group. Participants' Big Five personality data (Extraversion, Agreeableness, Conscientiousness, Emotional stability, Openness) were processed similarly. Previous experiences were divided into 'Low', 'Medium', and 'High' groups as previously stated.

Tables 4 and 5 compare Human-computer trust, Interpersonal trust, Specific trust behaviours and Overall trust mean scores with previous experience, trust propensity and Big Five personality levels. Previous experience levels were tested using the ANOVA test. Trust propensity and Big Five personality levels were tested using the T-test. The *p*-value indicated that only Specific trust behaviours would be negatively significantly affected by trust propensity while all the others would not. Participants with low trust propensity level had higher mean scores on specific trust behaviours. No significant effect on trust scores was found for either previous experience or Big Five personality.

Figure 6 displays the relationship between the experiment and control group in terms of Human-computer trust, Interpersonal trust, Specific trust behaviours and Overall trust. Each data point represented a pair of experiment + control group participants who used the same chatbot, which were all designed by the participants in the experiment group. The negative Pearson's *r* in figure 6 implied that there were negative correlations between the experiment group's trust scores and the control group's scores in all four trust dimensions. And the *p* value in figure 6b showed that there was a strong linear relationship between the results of the interpersonal trust scores in the experiment group and those in the control group. The higher the interpersonal trust score given by participants in the experiment group, the lower the interpersonal trust score given by participants who used the same chatbot in the control group. For human-computer trust and specific trust behaviour and overall trust dimensions, strong

Table 5: Results of independent sample T-tests, comparing different trust dimensions' score across trust propensity and Big Five personality types' level

Demographics	Trust dimensions	Group	N	Mean	SD	Shapiro-Wilk (p)	T-test (p)
Trust propensity	Human-computer trust	High	9	14.333	2.186	0.370	1.000
		Low	11	14.333	2.629	0.963	
	Interpersonal trust	High	9	20.889	5.159	0.662	0.744
		Low	11	20.091	5.504	0.201	
	Specific trust behaviours	High	9	5.111	2.147	0.363	0.015$*$
		Low	11	7.545	1.916	0.652	
	Overall trust	High	9	0.626	0.121	0.600	0.213
		Low	11	0.691	0.104	0.880	
Extraversion level	Human-computer trust	High	10	14.233	2.816	0.973	0.857
		Low	10	14.433	1.994	0.078	
	Interpersonal trust	High	10	20.600	4.551	0.451	0.902
		Low	10	20.300	6.075	0.074	
	Specific trust behaviours	High	10	6.700	1.494	0.138	0.644
		Low	10	6.200	3.011	0.886	
	Overall trust	High	10	0.669	0.104	0.451	0.775
		Low	10	0.654	0.128	0.745	
Agreeableness level	Human-computer trust	High	12	14.722	2.044	0.103	0.385
		Low	8	13.750	2.849	0.920	
	Interpersonal trust	High	12	21.167	5.844	0.106	0.468
		Low	8	19.375	4.274	0.972	
	Specific trust behaviours	High	12	6.500	2.714	0.964	0.910
		Low	8	6.375	1.768	0.926	
	Overall trust	High	12	0.678	0.126	0.389	0.458
		Low	8	0.638	0.096	0.313	
Conscientiousness level	Human-computer trust	High	12	14.278	2.620	0.045$*$	0.902
		Low	8	14.417	2.129	0.421	
	Interpersonal trust	High	12	20.250	5.987	0.090	0.840
		Low	8	20.750	4.200	0.519	
	Specific trust behaviours	High	12	6.083	2.539	0.610	0.403
		Low	8	7.000	2.000	0.975	
	Overall trust	High	12	0.647	0.133	0.901	0.508
		Low	8	0.683	0.081	0.441	
Emotional stability level	Human-computer trust	High	11	14.212	2.197	0.612	0.809
		Low	9	14.481	2.709	0.600	
	Interpersonal trust	High	11	20.364	5.887	0.452	0.937
		Low	9	20.556	4.640	0.362	
	Specific trust behaviours	High	11	7.000	2.828	0.612	0.253
		Low	9	5.778	1.394	0.062	
	Overall trust	High	11	0.675	0.131	0.435	0.566
		Low	9	0.645	0.093	0.612	
Openness level	Human-computer trust	High	13	14.179	2.686	0.556	0.704
		Low	7	14.619	1.820	0.606	
	Interpersonal trust	High	13	21.154	5.113	0.266	0.427
		Low	7	19.143	5.581	0.483	
	Specific trust behaviours	High	13	6.846	1.864	0.277	0.312
		Low	7	5.714	3.039	0.748	
	Overall trust	High	13	0.679	0.113	0.400	0.370
		Low	7	0.630	0.116	0.020$*$	



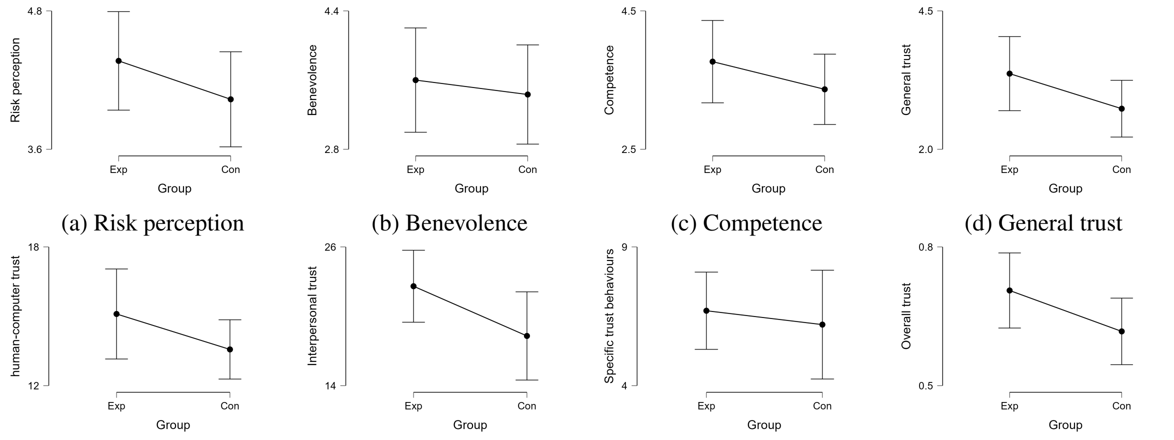
(a) Human-computer trust (n = 10, Pearson's r = -0.132, p = 0.716)

(b) Interpersonal trust (n = 10, Pearson's r = -0.647, p = 0.043*)

(c) Specific trust behaviours (n = 10, Pearson's r = -0.425, p = 0.221)

(d) Overall trust (n = 10, Pearson's r = -0.439, p = 0.204)

Figure 6: Scatter plot correlating different trust dimensions' score in experiment and control groups (Each data point represented a pair of experiment + control group participants who used the same chatbot, which were all designed by the participants in the experiment group)



(a) Risk perception

(b) Benevolence

(c) Competence

(d) General trust

(e) Human-computer trust

(f) Interpersonal trust

(g) Specific trust behaviours

(h) Overall trust

Figure 7: Descriptives plots of different trust dimensions' score in experiment and control group

linear relationships could not be found. In other words, there is no superiority or inferiority of chatbot design for these three trust dimensions, and it did not have a significant impact.

Table 6 and Figure 7 demonstrate the comparison of different trust dimensions and overall trust mean scores obtained from the experiment across the experiment and the control group. All T-test's p-values are > 0.05, from which it can be concluded that self-designing the anthropomorphic features of the chatbot did not statistically significantly affect the participants' trust in the chatbot. However, based on the figure and Cohen's d values, it can be seen that the effect sizes of some trust dimensions are relatively large, suggesting that the results appear to have practical meaning. Moreover, across all trust dimensions, the mean trust scores of participants in the experiment group were higher than those of the control group, indicating a consistent positive trend. Therefore, the main research statement of the paper can be summarised and answered as follows: although the self-designing chatbot's anthropomorphic features do not statistically significantly affect the participants' trust in the chatbot, they still seem to have a consistently positive effect.

In order to answer the second research question, Cohen's d value in Table 6 needs to be further analysed. It can be observed that Interpersonal trust (Cohen's d = 0.883) and Overall trust (Cohen's d = 0.825) had a large effect size. General trust came next, with an effect size that was close enough to the large criterion (Cohen's d = 0.760). Risk perception (Cohen's d = 0.568), Competence (Cohen's d = 0.517), and Human-computer trust (Cohen's d = 0.665) had medium effect sizes. This suggests that self-designing chatbots' anthropomorphic features have a strong positive effect on Human-Chatbot trust. It mainly affects Interpersonal trust and General trust, has a medium influence on Risk perception, Competence and Human-computer trust, and has a weak effect on Benevolence (Cohen's d = 0.203) and Specific trust behaviours (Cohen's d = 0.210). It is also worth mentioning that we could further observe that the T-test p values for both Interpersonal trust and Overall trust are close to the significance level. It strengthens our conjecture that self-designing anthropomorphic features may truly have an effect on Interpersonal trust and Overall trust, just not to the significance level due to the small sample size.

Table 6: Results of independent sample T-tests, comparing different trust dimensions' score in experiment and control group

Trust dimensions	Group	N	Mean	SD	Shapiro-Wilk (p)	T-test (t)	T-test (p)	T-test (Cohen's d)
Risk perception	Exp	10	4.367	0.597	0.106	1.270	0.220	0.568
	Con	10	4.033	0.576	0.288			
Benevolence	Exp	10	3.600	0.843	0.164	0.453	0.656	0.203
	Con	10	3.433	0.802	0.153			
Competence	Exp	10	3.767	0.832	0.046*	1.156	0.263	0.517
	Con	10	3.367	0.711	0.767			
General trust	Exp	10	3.367	0.936	0.818	1.699	0.106	0.760
	Con	10	2.733	0.717	0.252			
Human-computer trust	Exp	10	15.100	2.722	0.210	1.488	0.154	0.665
	Con	10	13.567	1.792	0.157			
Interpersonal trust	Exp	10	22.600	4.351	0.052	1.975	0.064	0.883
	Con	10	18.300	5.334	0.516			
Specific trust behaviours	Exp	10	6.700	1.947	0.046*	0.470	0.644	0.210
	Con	10	6.200	2.741	0.834			
Overall trust	Exp	10	0.706	0.114	0.539	1.845	0.082	0.825
	Con	10	0.617	0.101	0.235			

Table 7: Results of two-tailed paired samples T-tests (n=10), comparing different trust dimensions' score in experiment and control group

Measure 1		Measure 2	t	df	p	Cohen's d	SE Cohen's d
Exp_Human-computer trust	-	Con_Human-computer trust	1.405	9	0.194	0.444	0.499
Exp_Interpersonal trust	-	Con_Interpersonal trust	1.339	9	0.213	0.423	0.599
Exp_Specific trust behaviours	-	Con_Specific trust behaviours	0.397	9	0.700	0.126	0.536
Exp_Overall trust	-	Con_Overall trust	1.539	9	0.158	0.487	0.567

Furthermore, in order to eliminate the random variation in chatbot designs from the analysis, we ran two-tailed paired samples T-tests. Although the experiment is in a between-subjects design, the fact that each pair of experiment group and the control group participants shared a same chatbot makes the paired samples T-tests meaningful. The results are shown in Table 7. According to the p-values in the table, no statistically significant differences in mean values were found between the experiment and control groups. However, Cohen's d for Overall trust is close to the level of the medium effect size. This suggests that there may be a difference in trust scores between the experiment and control groups, regardless of how participants in the experiment group designed the chatbot anthropomorphic features. It is just not shown to be significant, likely because of the small sample size.

As mentioned before, Figure 7 shows a consistent positive trend. So we also conducted an upper-tailed paired T-test to check the significance in the positive direction. The results are shown in the Table 8. Although, as in the case of the two-tailed test, the p-values do not exhibit significant differences. Yet we found that the p-value for Overall trust was closer to the significance level. This implies a potential positive difference.

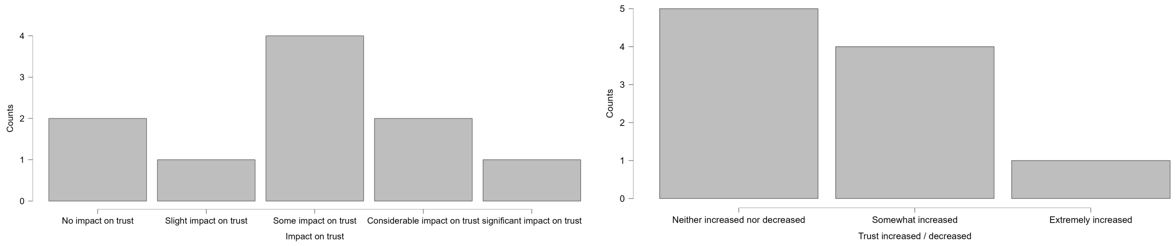
Qualitative analysis

For the experiment group, after the experiment, participants were asked whether they thought that participating in the self-design would affect trust in the chatbot as well as whether trust would increase or decrease and why. Their

Table 8: Results of upper-tailed paired samples T-tests (n=10), comparing mean different trust dimensions' score in experiment and control group

Measure 1		Measure 2	t	df	p	Cohen's d	SE Cohen's d
Exp_Human-computer trust	-	Con_Human-computer trust	1.405	9	0.097	0.444	0.499
Exp_Interpersonal trust	-	Con_Interpersonal trust	1.339	9	0.107	0.423	0.599
Exp_Specific trust behaviours	-	Con_Specific trust behaviours	0.397	9	0.350	0.126	0.536
Exp_Overall trust	-	Con_Overall trust	1.539	9	0.079	0.487	0.567

Note. For all tests, the alternative hypothesis specifies that Measure 1 is greater than Measure 2. For example, Exp_Human-computer trust is greater than Con_Human-computer trust .



(a) Whether affect trust

(b) Trust increased/ decreased

Figure 8: Distribution plots of participants' feedback in experiment group (n = 10)

Table 9: Descriptive Statistics of participants' feedback (Y = have an effect on trust, N = no effect on trust or thought it would neither increase nor decrease trust)

	Risk perception		Benevolence		Competence		General trust		human-computer trust		Interpersonal trust		Specific trust behaviours		Overall trust	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Valid	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Mean	4.600	4.133	3.733	3.467	4.133	3.400	3.533	3.200	16.000	14.200	22.400	22.800	7.200	6.200	0.734	0.678
Std. Deviation	0.435	0.691	0.863	0.901	0.380	1.038	0.960	0.989	2.068	3.220	4.336	4.868	2.490	1.304	0.112	0.120
Minimum	4.000	3.000	2.333	2.000	3.667	2.333	2.333	1.667	13.000	9.000	15.000	15.000	5.000	5.000	0.595	0.498
Maximum	5.000	4.667	4.667	4.333	4.667	4.667	5.000	4.333	18.667	17.000	26.000	28.000	11.000	8.000	0.861	0.837

results are shown in Figure 8 and Table 9. The main reasons for participants who thought there was an effect on trust and that it would increase trust were: (1) liking one of the anthropomorphic features of the chatbot and (2) finding the chatbot familiar. Participants who believed that there was no effect on trust or that it would neither increase nor decrease trust, their main reasons were: (1) the chatbot was just part of the game and would only give information, participants were just following the rules to play the game; (2) trust in the chatbot depended on their previous experiences and was not related to the game; (3) the most important thing of the chatbot was the functionality and not the appearance and other features.

From Table 9, it can be found that for participants who perceived an impact and those who perceived no effect on trust or thought it would neither increase nor decrease trust, their average interpersonal trust did not differ much, with the main differences being in human-computer trust and in trust behaviours. Meanwhile, except for interpersonal trust, participants who perceived an impact had higher mean scores on all other trust dimensions than those who perceived no impact or neither increase nor decrease. This may point to a real effect, although it cannot be statistically validated from our data.

For the control group, participants were asked after the experiment how much they liked the anthropomorphic features of the chatbot, whether they remembered the name of the chatbot, and if they could have participated in the self-design, whether they thought it would affect their trust in the chatbot. The results are shown in Figure 9. The majority of participants liked the appearance and voice of the chatbot assigned to them, and 50% counted as neither liking nor disliking the personality. 80% of participants felt that participation in the design of anthropomorphic features would affect trust in the chatbot. Only 4 of the 10 participants attempted to remember the chatbot's name, and only 2 actually remembered it. Interestingly, these two had the highest interpersonal trust (25 and 27) in the chatbot in the control group, and both believed that self-design had no or only a slight effect on trust in the chatbot.

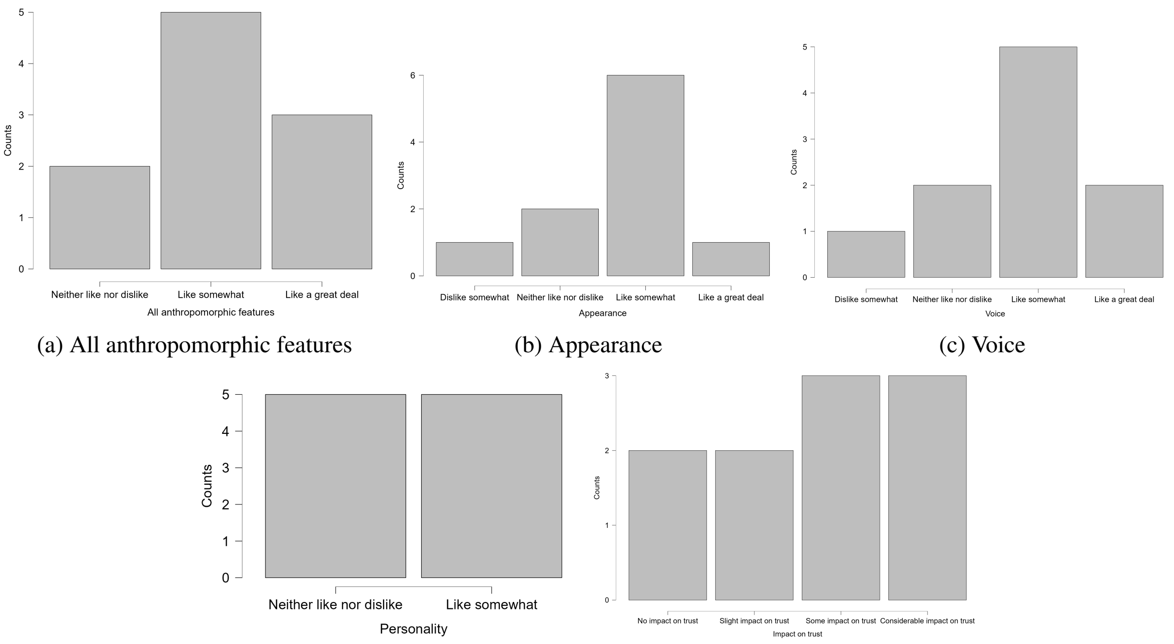
Discussion

In this section, we summarize the main and noteworthy findings of the experiment and attempt to give explanations. The limitations of the experiment and future work are also discussed.

Reflection on the results

From the above results analysis, it can be concluded that self-designing chatbots' anthropomorphic features have a positive effect on Human-Chatbot trust, which suggests its potential, though not statistically significant. This conclusion aligns with Wald et al. (2021), which states that the IKEA effect also holds true in terms of Human-Chatbot trust. The IKEA effect says that consumers are willing to pay more money for self-created products than similar products (Mochon, Norton, & Ariely, 2012).

Among the trust dimensions, the most positively influenced by self-designing anthropomorphic features in Human-Chatbot trust are interpersonal trust and general trust. This could be easily explained by the fact that, as demonstrated in the qualitative analysis, some of the participants in the experiment group felt that they had a sense of familiarity with the chatbots due to their self-design and liked their anthropomorphic features. This resulted in a higher interpersonal trust among participants in the experiment group than in the control group, and a



(a) All anthropomorphic features

(b) Appearance

(c) Voice

(d) Personality

(e) Whether affect trust

Figure 9: Distribution plots of participants' feedback in control group (n = 10)

higher trust as a whole.

The human-computer trust was moderately influenced by the self-designing chatbot anthropomorphic features, which is understandable. Since the participants in the experiment group were not involved in coding the chatbot, they were not entirely sure about the competence and safety of the chatbot. Their trust in safety and competence may derive more from using inworld.ai for design and therefore knowing the vendor and source of the chatbot. Benevolence trust was minimally affected by self-design, for the reason also mentioned in the qualitative analyses. Some of the participants in the experiment group thought that the chatbot was just a part of the game, just a device to provide information. These participants did not believe that it would act for the benefit of the participants.

The same weak influence of self-design was also observed in specific behaviours trust, which may be related to the influence of participants' trust propensity. In the quantitative analysis, it was shown that Specific trust behaviours would be significantly affected by trust propensity. This result is the same as the one obtained by Alarcon et al. (2018). However, unlike Alarcon et al. (2018)'s, this significance was negatively correlated, meaning that the lower the participant's trust propensity towards the chatbot, the higher their Specific trust behaviours score. After reviewing the original data, we discovered the possible reason for this. The Specific trust behaviour score consisted of 2 components, the number of times participants asked the chatbot for help and the number of times they followed the chatbot's advice. Some of the participants asked the chatbot for help multiple times, but only followed it a few times due to their lower propensity to trust. This made it difficult for them to solve the puzzle and thus they asked the chatbot for more help. This resulted in a high score regarding their specific behaviour. The opposite was also found for some participants with a high trust propensity. They asked only a few times to get the information they wanted and followed the advice to solve the puzzles, which made their scores seem low. For instance, there was one participant with a low trust propensity. He asked for help 6 times but only followed the advice 1 time. His score was 7. Another participant with a high trust propensity asked for help 4 times and followed the advice 2 times. He solved two puzzles but only scored 6, lower than the prior one. As concluded in the experiment observation subsection, just using trust behaviours to measure Human-Chatbot trust is not enough as it is greatly influenced by the special case.

In addition, when analysing the results, we came across some outcomes worth mentioning. The first one is the strong negative linear relationship between the interpersonal trust scores in the experiment group and those in the control group (Figure 6b). The higher the interpersonal trust scores given to the chatbot by participants in the experiment group, the lower the scores given by participants who used the same chatbot in the control group. This phenomenon may be explained using design theory. In our experiment, participants in the experiment group were equivalent to the designer and participants in the control group were equivalent to the consumer. When the designer integrated a lot of his or her personal preferences and cultural backgrounds into the design, rather than a universal design, he or she would have a higher degree of customisation of the chatbot and trusted the chatbot more interpersonally. But, correspondingly, a more specific chatbot will make it harder for consumers to find empathy. Consumers' inability to find elements they can relate to will make them trust less. The first half of this explanation is supported by the results of Lacroix, Wullenkord, and Eyssel (2022)'s study on robots, where they stated that affective trust in robots was significantly higher in the high degree of customisation condition than in the low degree of customisation, but cognitive trust was almost the same.

For consumers, as in the second part of the explanation, the situation might be different. People's reactions to design are innate, personal and cultural related. When designers and consumers are in different cultural backgrounds or have different personal characteristics, their tastes and understanding of the product may be very different (Crilly, Moultrie, & Clarkson, 2004). Thus, if the consumers in the control group have different cultural backgrounds or characteristics from the corresponding designers in the experiment group, this high degree of customisation would make it difficult for the consumers to approve of it, leading to lower interpersonal trust score.

This design theory also accounts for the human-computer trust dimension. In Figure 6, the Pearson's r for human-computer trust is the closest to 0, indicating that it has the weakest negative correlation between experiment and control group scores, compared to the other trust dimensions. In the design theory, preference, culture and degree of customisation mainly influence affective trust rather than cognitive trust, thus having a weak effect on human-computer trust.

Whether or not participants were finally successful in the game may also have an effect on the final trust scores, during the experiment period of this paper, a total of four participants were successes, three of which were from the experiment group. Using an independent t-test to analyse the different dimensions of trust scores across experiment success, it was found that both Overall trust and Special trust behaviours in the game success group were not normally distributed. The human-computer trust scores were normally distributed with a T-test $p = 0.018$, meaning that the results were significant. But the sample of the experiment is not enough for this result to be a solid conclusion. This is also the limitation of the experiment. Future experiments should be conducted with more participants and the success of the game should be included in the results analysis.

Known limitations

Based on what was mentioned in the reflections, the primary limitation of our experiment is the small sample size. With the current sample size, we can already observe consistent positive patterns and conclude that self-designing chatbot anthropomorphic features have great potential to affect Human-Chatbot trust. With more data from more participants, we could be able to verify the statistical significance better. A larger sample size would also help us reduce the bias caused by special cases and make the scoring of trust behaviours more convincing. In addition, a larger sample scale is also needed to study the effect of game success on the conclusions in this experiment.

Another limitation is the problem of appropriately quantifying specific trust behaviours. As previously mentioned, the number of times participants asked the chatbot and the number of times they followed the chatbot together comprised the trust behaviour score. However, sometimes, participants with a low trust propensity asked too much but did not follow, while some with a higher propensity asked less but followed the advice. This caused the latter to have lower scores than the former. These special cases made it difficult to quantify the behavioural scores reasonably. Giving different weights to help-asking and advice-following data might solve the problem, but further research is needed to determine the appropriate weights.

On the whole, the setting of the experiment was adequate for the research. Consistent positive effects can be concluded even with a small sample size and a short duration of the game. It could be interesting to extend it in various aspects though. In the following, we will suggest several research directions that might be fun to pursue. First is the time duration, our experiment lasted only 17 minutes, which included 2 minutes of familiarisation time with the chatbot and 15 minutes of game playing. This implies that the conclusions of our experiment only focus on people's initial trust in chatbots. The conclusions do not apply to long-term interactions with chatbots. Meanwhile, the experiment used an escape room game with a simple user interface and text-based interaction. More research should be conducted on how different game types, user interfaces, and interaction methods may affect Human-Chatbot trust.

The selection of chatbots and their anthropomorphic features is also noteworthy. In our experiments, we used the AI provided by inworld.ai, which uses a variety of large language models to provide services. Participants reflected their satisfaction with its overall capabilities, but still, one participant questioned the functionality of the chatbot after failing to find a clue. Differences in chatbot vendors, sources, and functionality are also worth adding as variables to the study. Furthermore, the options for chatbot anthropomorphic features are still limited. Our experiment allowed participants in the experiment group to design the chatbot's name, gender, appearance, voice, and personality, and the rest stayed in a uniform general setting. But the anthropomorphic features that can be designed are far more than that. The chatbot's dialogue style, actions during the narrative, etc. may have different effects on Human-Chatbot trust. This is also something that deserves more research in the future.

Future work

Besides the need for larger-scale experiments, further improvements to the Human-Chatbot trust model as well as to the measurements are also welcome. The present sub-dimension categories of human-computer trust are derived from Gulati et al. (2019), who categorised human-computer trust as Risk perception, Benevolence, Competence and General trust. However, a more detailed division could be achieved in future work, especially with the dimensions that have been shown to be medium or strong effects by the self-designing in this paper. For example, Risk perception can be categorised into distrust of chatbot vendors, insecurity about information leakage and feeling unsafe about AI.

Interpersonal trust is strongly influenced by self-designing anthropomorphic features and deserves to be highlighted. In our experiment, it was measured by six items from the Johnson-George and Swap (1982) Special Interpersonal Trust Scale. Future research could consider compiling the full Special Interpersonal Trust Scale, by modifying the other items to be more appropriate for Human-Chatbot trust. Or using other measurement scales that are more detailed.

The Special trust behaviour measurement, also mentioned in the results' reflection, needs to be optimised in advance for future study. Suitable scoring weights need to be figured out so that the results will not be heavily influenced by special cases and participant trust propensities. The time spent by participants seeking help from the chatbot and following advice should also be added for the future study's consideration.

Apart from the refinement of the trust model and measurements mentioned above, many directions of the experiment's design are worth expanding. First, the long-term impact of self-designing chatbot anthropomorphic features on Human-Chatbot trust is worthwhile to be researched. For example, we could consider designing a long-term task game that requires cooperation with a chatbot. The experiment asks participants to enter the game at least once a week and measure their trust in the chatbot. No restrictions are set on daily interactions. At the end of the experiment, record the number of times the participant plays the game and the change in his/her trust in the chatbot during a month. Initial trust, like the one we present in this paper, may determine whether or not people want to continue using the chatbot. Whereas, the frequency of people's use and retention of chatbots are supported by long-term trust. A study of long-term Human-Chatbot trust would not only tell the long-term impact of self-designing anthropomorphic features, but also the specific changing trends of how self-designing affects a particular trust dimension. This is helpful for the application of chatbots in a practical way.

In human-human trust building, playing a cooperative game is more effective than exchanging information (Depping et al., 2016), which is why we chose a cooperative escape room game for our experiment. However, it is not yet scientifically known whether cooperative games are really superior to information exchange in Human-Chatbot trust building and how much they differ. More research is needed in this area. Furthermore, different game genre comparisons could be a direction for expanding the research. Cooperative puzzle solving, cooperative mazes, cooperative business simulations, etc. Whether these game genres have a strong difference in people's trust in chatbots is an interesting topic to study. Such research could help those wishing to use chatbots for people's emotional support find the right design. Comparing different interaction methods such as text communication, voice communication, touch communication, etc. could also be a direction for further research. Nowadays most of the chatbots are still based on text communication and some of them have the option to allow voice communication. It is worth discussing which dimensions of trust are mainly affected by which interaction methods separately. A similar study had been done by Schreuter, van der Putten, and Lamers (2021). They compared the extent of people's conforming to conversational assistants across different modes of communication (text-based, robotic-voice and human-voice communication).

Differences in chatbots' vendor reputation may also be valuable study variables. For instance, conducting experiments using chatbots from well-known vendors, like OpenAi, versus lesser-known vendors while ensuring the basic functionality meets standards. Such a study could tell whether people's trust in chatbots differs depending on the source. As well, whether or not participants are informed of the chatbot's vendor could also be added as a variable to this research.

Finally, considering the limitations of our experiment on the choice of anthropomorphic features, more varied anthropomorphic features should be added to future studies. For example, allowing participants to self-design the chatbot's dialogue style or allowing participants to design a backstory for the chatbot. In fact, although we did not ask the participants to do so, one participant in our experiment spontaneously created a backstory for the chatbot he designed. These options that would make chatbots more anthropomorphic and customisable were not covered in our experiment, but would certainly be interesting to investigate as future directions.

Conclusion

Based on the existing literature on chatbots and trust, we compiled a Human-Chatbot trust model. In this model, Human-Chatbot trust is divided into six different dimensions: risk perception, benevolence, competence, general trust, specific interpersonal trust and specific trust behaviours. The first 4 of these can be summed up as human-computer trust. Based on this model, we pose 2 research questions: (1) Would people trust chatbots more if they designed the anthropomorphic features of the chatbot themselves? (2) What dimensions of Human-Chatbot trust are affected by self-designing the chatbot's anthropomorphic features?

To answer these two research questions, we designed a between-subjects experiment. Twenty participants were divided into experiment and control groups and took a 20-minute experiment in person. The experiment involved participants cooperating with a chatbot to play an escape room game. A few days before the experiment, participants in the experiment group were asked to design the anthropomorphic features (name, gender, appearance, voice, and personality) of a chatbot using inworld.ai, whereas participants in the control group did not do the designing and directly used the chatbot designed by the participants in the experiment group.

The results of the experiment answered the previously posed research question well. Although the self-designing chatbot's anthropomorphic features do not statistically significantly affect the participants' trust in the chatbot, they still appear to have positive effects with large effect sizes, which shows a strong potential to affect Human-Chatbot trust. Further research on a larger scale is needed to determine its statistical significance. Self-designing mainly affects Interpersonal trust and General trust, has a medium influence on Risk perception, Competence and Human-computer trust, and has a weak effect on Benevolence and Specific trust behaviours.

The research in this paper fills part of the knowledge gap on the effect of self-design of chatbot anthropomorphic features and contributes to future research on chatbot customisation.

References

Alarcon, G. M., Lyons, J. B., Christensen, J. C., Bowers, M. A., Klosterman, S. L., & Capiola, A. (2018). The role of propensity to trust and the five factor model across the trust process. *Journal of Research in Personality*, *75*, 69–82. doi: 10.1016/j.jrp.2018.05.006

Chandler, J., & Schwarz, N. (2010). Use does not wear ragged the fabric of friendship: Thinking of objects as alive makes people less willing to replace them. *Journal of Consumer Psychology*, *20*(2), 138–145. doi: 10.1016/j.jcps.2009.12.008

Cheng, X., Zhang, X., Cohen, J., & Mou, J. (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), 102940. doi: 10.1016/j.ipm.2022.102940

Crilly, N., Moultrie, J., & Clarkson, P. J. (2004). Seeing things: consumer response to the visual domain in product design. *Design Studies*, *25*(6), 547–577. doi: 10.1016/j.destud.2004.03.001

Degachi, C., Tielman, M. L., & Al Owayyed, M. (2023). Trust and perceived control in burnout support chatbots. In *Extended abstracts of the 2023 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3544549.3585780

Depping, A. E., Mandryk, R. L., Johanson, C., Bowey, J. T., & Thomson, S. C. (2016). Trust me: Social games are better than social icebreakers at building trust. In *Proceedings of the 2016 annual symposium on computer-human interaction in play* (pp. 116–129). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2967934.2968097

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology. Applied*, *22*(3), 331–349. doi: 10.1037/xap0000092

Ehrhart, M. G., Ehrhart, K. H., Roesch, S. C., Chung-Herrera, B. G., Nadler, K., & Bradshaw, K. (2009). Testing the latent factor structure and construct validity of the ten-item personality inventory. *Personality and Individual Differences*, *47*(8), 900–905. doi: 10.1016/j.paid.2009.07.012

Fahr, R., & Irlenbusch, B. (2008). Identifying personality traits to enhance trust between organisations: An experimental approach. *Managerial and Decision Economics*, *29*(6), 469–487. doi: 10.1002/mde.1415

Furumo, K., de Pillis, E., & Green, D. (2009). Personality influences trust differently in virtual and face-to-face teams. *International Journal of Human Resources Development and Management*, *9*(1), 36–58. doi: 10.1504/IJHRDM.2009.021554

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, *38*(10), 1004–1015. doi: 10.1080/0144929X.2019.1656779

Guo, Y., Wang, J., Wu, R., Li, Z., & Sun, L. (2022). Designing for trust: A set of design principles to increase trust in chatbot. *CCF Transactions on Pervasive Computing and Interaction (Online)*, *4*(4), 474–481. doi: 10.1007/s42486-022-00106-5

Guthrie, S. (1993). *Faces in the clouds: A new theory of religion.* Oxford University Press.

Hale, J., Payne, M. E., Taylor, K. M., Paoletti, D., & De C Hamilton, A. F. (2018). The virtual maze: A behavioural tool for measuring trust. *Quarterly Journal of Experimental Psychology (2006)*, *71*(4), 989–1008. doi: 10.1080/17470218.2017.1307865

Handy, B. P. (2018). Building interpersonal trust through digital games. *ProQuest Dissertations and Theses*, 53.

Johnson-George, C., & Swap, W. C. (1982). Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of Personality and Social Psychology*, *43*(6), 1306–1317. doi: 10.1037/0022-3514.43.6.1306

Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? a systematic review. *Telematics and Informatics*, *77*, 101925. doi: 10.1016/j.tele.2022.101925

Komiak, S. Y. X., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, *30*(4), 941–960. doi: 10.2307/25148760

Lacroix, D., Wullenkord, R., & Eyssel, F. (2022). I designed it, so i trust it: The influence of customization on psychological ownership and trust toward robots. In *Social robotics* (pp. 601–614). Springer Nature Switzerland. (in press) doi: 10.1007/978-3-031-24670-8$_{5}$3

Lee, K. M., Peng, W., Jin, S.-A. A., & Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication*, *56*(4), 754–772. doi: 10.1111/j.1460-2466.2006.00318.x

Lin, J., Cronjé, I., Käthner, P., Pauli, P., & Latoschik, M. E. (2023, May). Measuring interpersonal trust towards virtual humans with a virtual maze paradigm. *IEEE Transactions on Visualization and Computer Graphics*, *29*(5), 2401–2411. doi: 10.1109/TVCG.2023.3247095

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. doi: 10.2307/258792

Mochon, D., Norton, M. I., & Ariely, D. (2012). Bolstering and restoring feelings of competence via the ikea effect. *International Journal of Research in Marketing*, *29*(4), 363–369. doi: 10.1016/j.ijresmar.2012.05.001

Moradinezhad, R., & Solovey, E. T. (2021). Investigating trust in interaction with inconsistent embodied virtual agents. *International Journal of Social Robotics*, *13*(8), 2103–2118. doi: 10.1007/s12369-021-00747-z

Morita, P. P., & Burns, C. M. (2014). Understanding "interpersonal trust" from a human factors perspective: insights from situation awareness and the lens model. *Theoretical Issues in Ergonomics Science*, *15*(1), 88–110. doi: 10.1080/1463922X.2012.691184

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, *43*(2), 223–239. doi: 10.1006/ijhc.1995.1042

Nass, C., Steuer, J., & Tauber, E. (1994). Computers are social actors. In *Chi '94 proceedings of the sigchi conference on human factors in computing systems* (pp. 73–78). Boston, MA. doi: 10.1145/191666.191703

Neave, N., Jackson, R., Saxton, T., & Hönekopp, J. (2015). The influence of anthropomorphic tendencies on human hoarding behaviours. *Personality and Individual Differences*, *72*, 214–219. doi: 10.1016/j.paid.2014.08.041

Niewiadomski, R., Demeure, V., & Pelachaud, C. (2010). Warmth, competence, believability and virtual agents. In *Iva* (pp. 272–285). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-15892-6$_{2}$9

Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2019). An initial model of trust in chatbots for customer service—findings from a questionnaire study. *Interacting with Computers*, *31*(3), 317–335. doi: 10.1093/iwc/iwz022

Pesonen, J. A. (2021). 'are you ok?' students' trust in a chatbot providing support opportunities. In *Learning and collaboration technologies: Games and virtual environments for learning* (pp. 199–215). Springer International Publishing. doi: 10.1007/978-3-030-77943-6$_{1}$3

Riedl, R. (2022). Is trust in artificial intelligence systems related to user personality? review of empirical evidence and future research directions. *Electronic Markets*, *32*(4), 2021–2051. doi: 10.1007/s12525-022-00594-4

Roy, R., & Naidoo, V. (2021). Enhancing chatbot effectiveness: The role of anthropomorphic conversational styles and time orientation. *Journal of Business Research*, *126*, 23–34. doi: 10.1016/j.jbusres.2020.12.051

Schillaci, C. E., de Cosmo, L. M., Piper, L., Nicotra, M., & Guido, G. (2024). Anthropomorphic chatbots for future healthcare services: Effects of personality, gender, and roles on source credibility, user satisfaction, and intention to use. *Technological Forecasting & Social Change*, *199*, 123025. doi: 10.1016/j.techfore.2023.123025

Schreuter, D., van der Putten, P., & Lamers, M. H. (2021). Trust me on this one: Conforming to conversational assistants. *Minds and Machines (Dordrecht)*, *31*(4), 535–562.

Seitz, L., Bekmeier-Feuerhahn, S., & Gohil, K. (2022). Can we trust a chatbot like a physician? a qualitative study on understanding the emergence of trust toward diagnostic chatbots. *International Journal of Human-Computer Studies*, *165*, 102848. doi: https://doi.org/10.1016/j.ijhcs.2022.102848

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion - a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, *149*, 102601. doi: 10.1016/j.ijhcs.2021.102601

Syrdal, D. S., Dautenhahn, K., Woods, S. N., Walters, M. L., & Koay, K. L. (2007). Looking good? appearance preferences and robot personality inferences at zero acquaintance. In *Aaai spring symposium: Multidisciplinary collaboration for socially assistive robotics.*

Wald, R., Heijselaar, E., & Bosse, T. (2021). Make your own: The potential of chatbot customization for the development of user trust. In *Adjunct proceedings of the 29th acm conference on user modeling, adaptation and personalization* (p. 382–387). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3450614.3463600

Wang, W., Qiu, L., Kim, D., & Benbasat, I. (2016). Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. *Decision Support Systems*, *86*, 48–60. doi: 10.1016/j.dss.2016.03.007

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. doi: 10.1016/j.jesp.2014.01.005

Xiao, J., Stasko, J. T., & Catrambone, R. (2007). The role of choice and customization on users' interaction with embodied conversational agents: effects on perception and performance. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Yamagishi, T., & Yamagishi, M. (1994, June). Trust and commitment in the united states and japan. *Motivation and Emotion*, *18*(2), 129–166. doi: 10.1007/bf02249397

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics - Association for Computational Linguistics*, *46*(1), 53–93. doi: 10.1162/coli_a_00368

Appendix

Trust propensity questionnaire comparison

Interpersonal trust propensity	Human-Chatbots trust propensity
Most people are basically honest.	Most chatbots are basically honest.
Most people are trustworthy.	Most chatbots are trustworthy.
Most people are basically good and kind.	Most chatbots are basically good and kind.
Most people are trustful of others.	Most people are trustful of chatbots.
I am trustful.	I am trustful.
Most people will respond in kind when they are trusted by others.	Most chatbots will respond in kind when they are trusted by people. (respond in kind in this case means you react to something that someone has done to you by doing the same thing to them.)

Table 10: Trust propensity questionnaire comparison