Chatbots as Advisers: the Effects of Response Variability and Reply Suggestion Buttons

Federico Milana
federico.milana.18@ucl.ac.uk
UCL Interaction Centre
London, United Kingdom

Enrico Costanza
e.costanza@ucl.ac.uk
UCL Interaction Centre
London, United Kingdom

Joel Fischer
joel.fischer@nottingham.ac.uk
University of Nottingham
Nottingham, United Kingdom

Abstract

As chatbots gain popularity across a variety of applications, from investment to health, they employ an increasing number of features that can influence the perception of the system. Since chatbots often provide advice or guidance, we ask: do these aspects affect the user's decision to follow their advice? We focus on two chatbot features that can influence user perception: 1) response variability in answers and delays and 2) reply suggestion buttons. We report on a between-subject study where participants made investment decisions on a simulated social trading platform by interacting with a chatbot providing advice. Performance-based study incentives made the consequences of following the advice tangible to participants. We measured how often and to what extent participants followed the chatbot's advice compared to an alternative source of information. Results indicate that both response variability and reply suggestion buttons significantly increased the inclination to follow the advice of the chatbot.

CCS Concepts

• **Human-centered computing** → **Natural language interfaces**; **User studies**; • **Information systems**;

Keywords

Chatbot, interface design, user behaviour, social trading, online study

ACM Reference Format:
Federico Milana, Enrico Costanza, and Joel Fischer. 2023. Chatbots as Advisers: the Effects of Response Variability and Reply Suggestion Buttons. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3571884.3597132

1 Introduction

Enabled by deep learning and large training datasets, advances in natural language processing have made chatbots increasingly popular in a wide variety of application areas. Advantages of chatbots include instant, 24-hour responses and savings on personnel, which can be critical where resources are scarce. Examples range from customer service to information acquisition to health (with a variety of "chatbot doctors" already on the market). In most of these applications, the user's inclination to follow the guidance and advice of the chatbot is a crucial component to take into consideration when designing these services and measuring their success. As natural language processing evolves and chatbots become more widespread, we investigate whether the inclination to follow advice of a chatbot is influenced by two novel features that have been widely adopted recently: response variability and reply suggestion buttons.

The effects of answer variability on the perception of conversational agents have been observed by Xuetao et al., who demonstrated that the biggest drawback of chatbots, the lack of believability, could be bypassed by varying the chatbot's answers [41]. Our study considers response time as an additional aspect that has also been revealed to influence the perception of chatbots [21, 26]. In "response variability", the study includes both answer variability and delay variability as design aspects that can introduce cognitive biases.

Reply suggestion buttons have also been recently integrated into chatbot UIs. Once the user receives a message, a small number of buttons is displayed, each representing a contextually relevant answer. Users can press one of the buttons as an alternative way of sending a message. The buttons increase usability and constrain user input, preventing typos and making it easier for the chatbot to process a message. Based on the relevant literature, we expect these to contribute to a greater perception of a competent chatbot. We investigate the effects of reply suggestion buttons on user behaviour and whether their presence interacts with other features that can influence how the chatbot is perceived.

This paper reports on a between-subject online study designed around a simulated social trading platform scenario. Social trading platforms, such as eToro[1] and ZuluTrade[2], are financial investment platforms where investors can simply copy the investments of other traders by "following" their portfolios. This scenario is convenient for our study for multiple reasons. Firstly, it involves explicit investment choices that can be influenced by information provided by the chatbot and the perception of the agent. By tracking the participants' investments, we can measure how often and to what extent they follow the advice of the chatbot compared to an alternative source. Additionally, the monetary aspect facilitates the implementation of financial study incentives, which follows similar research methods in behavioural economics recently adopted into HCI [39]. Finally, compared to a traditional financial investment scenario, social trading does not require and is not influenced by specialist knowledge, allowing us to recruit participants from the general population.

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

[1] https://www.etoro.com/
[2] https://www.zulutrade.com/

Our results indicate that both varied, dynamically delayed answers and reply suggestion buttons have a significant effect on user perception and behaviour, the key *contribution* of this paper. Participants followed investment advice more frequently when the chatbot phrased replies differently each time with delays based on text length. Participants also followed the advice more frequently when buttons were available.

2 Related Work

Herein we review related work at the intersection of our contribution on chatbots, response variability, reply suggestion buttons and social trading.

2.1 Chatbots

The perception of chatbots has been proven to influence attitudes, satisfaction and emotional connection between consumer and company [3]; aspects of pressing concern as the interface between companies and consumers gradually evolves to become technology dominant rather than human-driven [30]. Meanwhile, recent work has highlighted the potential for relatively simple UI variations to change user perception of AI systems [20, 25] revealing that, as the complexity (perceived or real) of interactive systems increases, cognitive biases may influence users understanding of them.

The literature points to numerous effects of social cues, such as informal or personalised speech, taking place on the interaction with a chatbot. For example, Liu et al. reported that health advice given with expressions of sympathy and empathy is favoured by users, especially by those who are initially sceptical of the competence of conversational agents [32]. In fact, since computer-mediated communication can display emotion as well as, or better than, face-to-face communication [15], users expect chatbots to sustain conversations of emotional nature, with over 40% of customer service requests to chatbots being of this kind [40]. The use of social cognitive abilities from chatbots has been found to have a significant effect on perceived believability, which, in turn, fosters an emotional connection between consumer and company [3].

In recent work, Gao et al. [18] proposed a computational approach to extracting features and training models that predict chatbots' popularity and performance. Readability and representation of responses emerged as the most important features for chatbot popularity: users prefer communicating with chatbots whose responses contain more links and HTML structures. However, whether this is due to their perception of the system or simply due to usability factors (links can help users access external resources more easily) was not explored. Prior work on reply suggestion buttons in chatbots reported that they consistently improve response efficiency, reducing response time by 12-35% and keystrokes by 33-60% [19]. According to a recent study on the determinants for the adoption of intelligent agents across the medical imaging workflow [6], ease of use is a significant antecedent of the user's performance expectancy of AI systems in general. In this context, we aim to evaluate the effect of reply suggestion buttons on performance expectancy based on the increased efficiency introduced by the buttons.

Since a considerable amount of chatbots is implemented in healthcare, investment and customer service [12, 17], decision-making is typically required by the user at some point in the interaction. In electronic commerce, we know that the consumer's decision-making process is heavily influenced by the disposition to trust, privacy concerns, information quality and company reputation [28]. The valence framework proposed by Tarpey and Peter [36], consistent with Lewin's [31] and Bilkey's [4] models, suggests that these factors contribute to the formation of positive and negative attributes for each product. It is the net sum of these attributes that determines the consumer's decisions. This equally applies to an investment environment, where the opinion of a conversational agent creates an additional attribute to each choice alternative that the user takes into consideration. Whether the weight, or valence, of this attribute is greater if the chatbot features variable responses and reply suggestion buttons has not been investigated before.

2.1.1 Response Variability The literature on chatbot language design emphasises a lack of consistency in style [33] and recent work has focused on identifying methods of consciously tailoring register for specific contexts [10]. However, despite influencing credibility, trust and persuasiveness [9], register-specific language alone fails to address the technical limitations of closed-domain, task-oriented chatbots. In fact, since tasks and sub-tasks are performed repeatedly (providing financial advice, asking for symptoms or purchasing a product), sentence structure must vary to avoid repeating identical or extremely similar messages. In a study where 21 subjects interacted with six agents with different degrees of variability, Xuetao et al. reveals that agents with increased variability in their answers are perceived as more believable, cooperative and satisfying to use [41].

Meanwhile, the literature on response delays of conversational agents suggests that response time has a significant effect on how a chatbot is perceived and that instant replies fail to give users the feeling of a natural conversation [2, 26]. To avoid this, modern chatbots intentionally delay responses to simulate a person typing despite calculating the appropriate reply almost instantly [29]. Delays not only increase the perception of social presence but also lead to greater satisfaction with the overall chatbot interaction [21]. Research also suggests that users expect delay times to vary depending on the content of both the request and response [34]. Therefore, responses should vary and appear with a dynamic delay to replicate a successful implementation of a believable chatbot.

2.1.2 Reply Suggestion Buttons The literature on reply suggestions is somewhat lacking, given the novelty of the feature. However, research made on Google Inbox's Smart Reply can be considered in this context. In particular, the investigation of this feature was set to discover whether it was possible to assist users with composing short messages by suggesting brief responses when appropriate [27]. Compared to the implementation inside a conversational agent, where reply suggestions can be generated according to the messages sent by the chatbot, Smart Replies for emails require a higher degree of response prediction accuracy. The generation of reply suggestions in chatbots requires no neural natural language understanding models, which removes a significant challenge in development [24]. For this reason, while the Smart Reply system was responsible for assisting with only 10% of email replies for Inbox on mobile in 2016 [27], in a recent study by Gao et al. [19], users have been found to adopt suggestions without any changes in 44%–68% of the cases when sending messages to a chatbot.

While the presence of suggestion buttons improves response quality sometimes, and only slightly, response efficiency is greatly affected in question-answering conversations because of two reasons: suggestions can provide critical relevant information, and users can reuse the suggestions' text [19]. This holds true regardless of the performance of the chatbot.

A recent study on the effects of visual and conversational cues on the perception of chatbots reveals that high message interactivity facilitates the feeling of interacting with another person [22]. In a similar note, work by Sundar et al. suggests that the longer the back-and-forth exchange between two humans, the stronger the feeling of the other's presence [38]. From this, it follows that reply suggestion buttons reduce perceived humanness as they effectively shorten the exchange by increasing the usability of the interface. However, recent work from García et al. [20] reveals that visual animation cues change people's perception of how well an intelligent system performs its task regardless of how human-like the animations appear to be. Therefore, reply suggestion buttons might still inadvertently influence the perception of competence despite decreasing the believability of the agent.

2.2 Social Trading

Social trading platforms provide access to an innovative type of delegated portfolio management by allowing investors to copy the investment strategies of other traders [16]. Prior work suggested that while social trades outperform individual trades, the social reputation of the top traders represented as their number of followers is not entirely determined by their performance [35]. In our study it was desirable to minimise the number of external factors that the user is required to analyse before taking a decision to avoid the interpretation of a user's action becoming unnecessarily complicated. For instance, if the agent's advice is disregarded, how much weight did the user assign to the trader's number of followers compared to the trader's performance? The necessary compromise of removing social feedback implies an increased asymmetry in the information held by users and traders. While social trading platforms aim to reduce information asymmetry by publishing a full single transaction history and standardised real-time track records for each portfolio [16], this is likely to detriment an analysis of data produced by our study.

An issue then arises as to how much symmetry is required between information held by the user and that held by traders in order to simulate a similar platform adequately. For example, there is strong evidence that trader communication impacts investment decisions of followers, despite comments not containing informational value [1]. In order to simplify the interpretation of actions from the user, the study considered this a valid reason to exclude communication from traders in the simulated environment. On the other hand, since these platforms claim to offer a variety of parameters to choose from, such as gain, risk score and location, which are advertised as primary features[3], forbidding access to this information would have affected the essence of social trading excessively.

[3]https://www.etoro.com/discover/people; https://www.zulutrade.com/

3 Study Design

Informed by the concepts of the perception of chatbots and social trading explored in the literature, we designed a between-groups study to measure the influence of chatbot features response variability and reply suggestion buttons on user decision-making, particularly on advice-following. We also assessed whether the effect of response variability interacts with the effect of reply suggestion buttons through a 2x2 factorial study design.

We designed our study around a simulated social trading platform. To simulate a realistic interaction with a chatbot, where the user might not focus their entire attention on the chat, participants could switch between two tasks freely: 1) social trading and 2) image tagging. The inclusion of a secondary task also provided the means to distract participants from the main issue being examined.

Both tasks were financially rewarded but subject to a time limit. In this context, if participants spent time on the trading task, they lost the chance of earning money through the image-tagging task, with an associated opportunity cost caused by the time limit. This study design was adapted from prior work studying user interaction with AI and autonomous systems [39].

Participants were allocated a budget of 1000 virtual pounds at the start of the study. The budget could be invested by following portfolios and increased by performing the image-tagging task (+20 virtual pounds per image). Participants were compensated £5 for their time, plus a bonus reward: £1 for every 200 virtual pounds gained above the initial 1000 virtual pounds.

3.1 Materials: Simulated Social Trading Platform

A simplified social trading platform was designed and implemented for the study. We release the code as open source [4]. As illustrated in Figure 1, the main view includes a set of portfolios, a chatbot, a news feed, an image tagging interface, and a study information bar, each described in the following subsections. The platform was implemented using a combination of HTML5 and the Django web framework and hosted on a publicly accessible web server. The chatbot was built on Rasa, an open-source machine-learning framework for text- and voice-based assistants.

3.1.1 Portfolios Portfolios provide users with the decision to copy trading strategies of experienced traders. By following a portfolio, the user decides to invest a portion of their capital so that its value can change over time depending on the performance of the associated trader. Users can follow or unfollow a portfolio, and add or withdraw investment amounts from an already followed portfolio.

The platform includes 10 portfolios. Trader names were fictional and generated to be uncommon to avoid familiarity bias. The performance of each portfolio is calculated based on a normally distributed random variable with a mean of 0.00 and standard deviation equal to the portfolio risk score. The risk score is a value between 1 and 10, randomly assigned to each portfolio and kept constant throughout the study. Performance changes were limited to between -99.99% and +99.99% to maintain a minimal degree of realism.

While portfolios on real social trading platforms change continuously and users can start or stop following them at any time,

[4]https://github.com/fmilana/chatbot

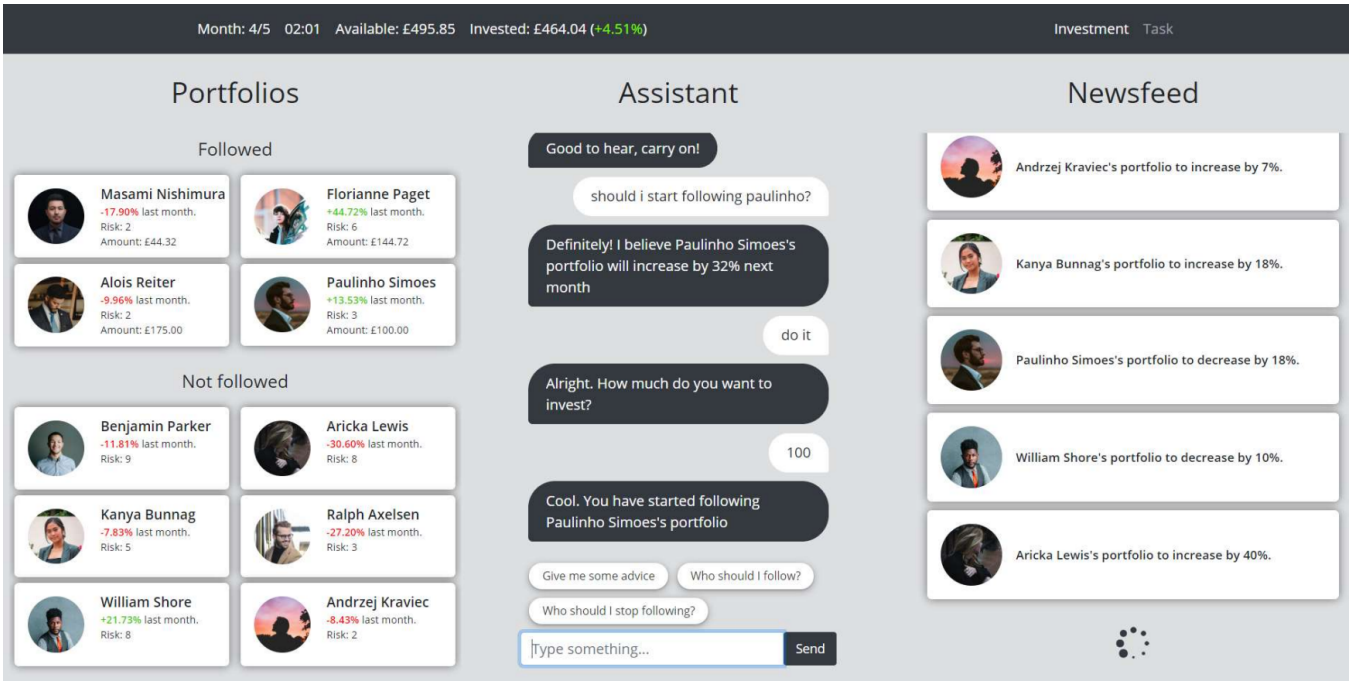

Figure 1: Main view of the web application. On the left, the list of portfolios, each with the fictional name and profile picture of the traders, the recent performance and risk level. In the middle, the chatbot assistant panel, shown here for the condition with the reply suggestion buttons (just above the text input box). On the right, the news feed periodically showing predicted changes to the portfolios, each including the related profile picture.

changes can take place only at discrete intervals in our simulation. We call these intervals virtual months, with each virtual month corresponding to 4 minutes of study time. In total, there were 5 virtual months within the study duration of 20 minutes. The discrete approach was taken to simplify the simulation and, more importantly, to avoid time pressure on the participants (they would otherwise be required to perform actions as fast as possible in order to avoid immediate losses).

Two gaussian random variables with mean 0 and variance 1 are generated for each portfolio each virtual month. The investment advice provided by the chatbot is based on one variable, while the advice provided by the news feed is based on the other. At the beginning of each month, one variable is randomly selected with equal probability to generate the actual change for each portfolio. This way, each source of advice will be correct on average 50% of the time and users should not be able to trust or mistrust either source based on observed performance. Since the two variables are generated independently from each other, they may both have positive or negative values.

3.1.2 Chatbot The chatbot is displayed on the central column of the user interface. It offers investment advice and executes actions on portfolios on behalf of the user. As previously mentioned, the chatbot has access to one random variable for each portfolio, which may or may not be accurate. Based on this information, the chatbot can provide advice about which portfolio to start following (the one with the highest predicted gain) or to stop following (the one with the highest predicted loss), as well as advice about any individual portfolio specified by the user.

In addition to asking for advice, users need to instruct the agent to perform any operations on the portfolios. The chatbot is the only way for users to interact with the trading environment. The purpose was to maximise the interaction with the chatbot; the focus of our study. While such an arrangement may well influence participants to pay more attention to the chat compared to the news feed, this would affect all conditions equally and should not have any effect on the comparative results.

At the beginning of the study, the chatbot informs users about its functions (i.e. that they can request advice and perform actions on portfolios). If a user does not interact with the agent for more than 45 seconds at any point during the study, it will spontaneously offer advice regarding which portfolio to follow or stop following in order to increase the chances of interaction.

3.1.3 News Feed The news feed is displayed on the right-hand side of the interface and shows a new post at random intervals between 15 and 25 seconds to simulate a realistic feed. Each post contains a prediction of a random portfolio's future gain or loss. It is made clear to the user that posts have a different source of information than the chatbot and that the predictions between the two might not coincide.

The primary role of the feed is to provide a source of advice alternative to the chatbot. While real social trading platforms disclose

exhaustive data on past performance for each trader, the interpretation of data relies excessively on the user's knowledge of trading intricacies. In this implementation, as the news feed is producing simple portfolio predictions, participants can simply read posts from a feed without any specialist knowledge.

3.1.4 Secondary Task: Image Tagging An image-tagging activity in order to distract participants from the primary social trading task was integrated into the platform, illustrated in Figure 2. Image tagging was opened by clicking on the "Task" tab and replaced the portfolios on the left, leaving the chatbot and the news feed visible to the user. In this view, participants were asked to guess 3 out of 6 possible tags describing an image. After guessing 3 tags successfully, 20 virtual pounds was added to their available balance. There were 60 available images for participants to tag, skip or re-attempt later in the study. Pictures and tags were taken from Flickr[5].

3.1.5 Study Information Bar At the top of the interface, participants could see information about their current progress in the study. The information bar includes the current virtual month, the time left until the end of the month, the amount of virtual money available and the amount invested.

3.2 Procedure

The study was run online. After providing informed consent, participants read a tutorial explaining the study. The main section of the study lasted 20 minutes, corresponding to 5 virtual months. Participants were notified of the end of each month by a pop-up dialogue box.

3.3 Independent Variables and Conditions

The study included 4 conditions in a 2x2 factorial design with 2 independent variables: "response variability" and "reply suggestion buttons". The first variable, response variability, contains 2 levels:

- **RV** — answers are varied and dynamically delayed.
- **NRV** — answers are not varied nor dynamically delayed.

The second variable "reply suggestion buttons" also includes 2 levels:

- **RS** — reply suggestions buttons are available.
- **NRS** — no reply suggestion buttons.

In the RV level, each message from the chatbot was randomly chosen from a list of 10 utterances with the same meaning (Table 1). In contrast, the chatbot in the NRV level used the same utterances for each message sent. Chatbot messages in the RV level were delayed depending on their length, by 45ms per character (with a minimum of 1200ms). Chatbot messages in the NRV level were displayed after a mere 1000ms (the time required to process the input). In both levels, a typing indicator was shown whenever a message was loading. Each participant only experienced one type of chatbot and the study followed a between-groups design.

3.4 Participants

A total of 64 participants, 16 per condition, were recruited from Prolific[6], an online crowdsourcing platform for participant recruitment. All were fluent in English. The only criteria set for participant recruitment was minimum age: 18. Participants demographics are summarized in Table 2.

3.5 Analysis

We were particularly interested in the following dependent variables:

3.5.1 Follow Ratio To calculate how often participants followed the advice of the chatbot compared to the predictions of the news feed, we restricted the analysis to situations where the chatbot and the news feed offered contradictory advice, i.e. one showing a positive variability and the other a negative variability. The follow ratio was calculated as the number of times in which the chatbot's advice was followed divided by the total number of times in which the predictions of the two sources were contradictory. That is, out of all the actions performed by the participant (follow, unfollow, add or withdraw) where the chatbot predicted a positive change and the news feed a negative change (or vice-versa), what percentage of these were in accordance with the chatbot's advice?

3.5.2 Additional Dependent Variables Additional dependent variables included: the average action size (the amount of virtual pounds involved) following the chatbot's advice in situations of contradictory predictions, the number of messages sent, the number of actions performed, the number of tag attempts and the number of correct tag attempts.

4 Results

4.1 General Descriptive Statistics

On average, each participant sent 59.98 messages to the chatbot (SD = 22.89). Of these, 22.77 (SD = 9.10) were actions on the portfolios. Each participant attempted 78.30 (SD = 67.43) image tags on average, 35.08 (SD = 31.57) of which were correct. The average final total balance was 1478.00 virtual pounds (SD = 502.40), corresponding to an average bonus of £2.39.

Each participant experienced an average of 11.75 (SD = 5.65) contradictory situations, where the chatbot and news feed advice was in opposite directions. In these occasions, each participant followed the advice of the chatbot 7.08 times (SD = 4.22) and the predictions of the news feed 4.67 times (SD = 3.63). The average chatbot follow ratio across all conditions was 0.68 (SD = 0.22). See Figure 3 for the distribution of the follow ratio across conditions. The size of trading actions following the chatbot's advice in these situations was £159.10 (SD = £131.10), while the size of actions following the prediction of the news feed was £183.60 (SD = £192.50).

4.2 Inferential Statistics

A two-way ANOVA revealed a statistically significant effect of both response variability (p = 0.009, F = 7.387) and reply suggestions (p < 0.001, F = 14.018) on the follow ratio. For participants exposed

[5]https://www.flickr.com/
[6]https://www.prolific.co/

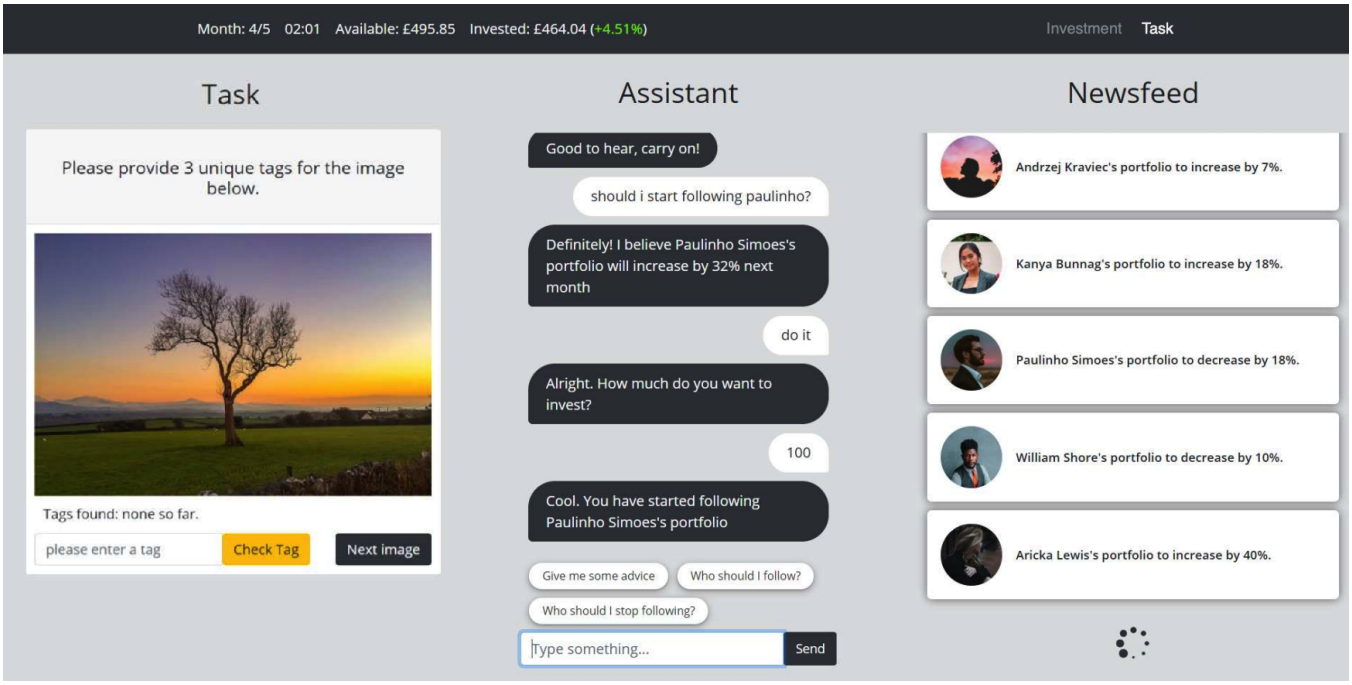

Figure 2: The view of the web application, with the image-tagging panel replacing the list of portfolios on the left. The central and right panels are identical to those in Figure 1.

Chatbot utterance
"You're doing great! I don't think you should follow or unfollow anyone else this month"
"I don't think there is anyone you should follow or unfollow currently"
"That's it, you're doing great! I don't think there is anyone else you should follow or unfollow"
"I can't think of anyone else you should follow or unfollow at the moment. You're doing great!"
"I don't think you should follow or unfollow anyone else at the moment"
"That's it! I don't think there is any other portfolio you should follow or unfollow this month"
"You're doing great! I don't think there is anyone else you should follow or unfollow at the moment"
"I don't think there is anyone else you should start or stop following at the moment"
"You're doing great! I can't think of anyone else to follow or unfollow"
"I don't think you should follow or unfollow anyone else this month"

Table 1: Variability of the message sent when the chatbot had no advice left for the month

Age	$AVG = 26.7, SD = 9.6$
Sex	43 Male, 21 Female
Nationality	25 United Kingdom, 10 Poland, 8 Portugal, 21 Other
Highest education	32 Secondary school, 20 Undergraduate, 8 Graduate, 2 No formal qualifications, 2 Other
Employment	21 Unemployed, 20 Full-time, 12 Part-time, 5 Not in paid work, 6 Other
Fluent languages	44 English, 8 Polish, 5 Portuguese, 11 Other
Literacy difficulties	60 No, 4 Yes

Table 2: Summary of the participants demographic information (Complete dataset referenced in the acknowledgements)

to response variability, the follow ratio was on average 0.75 (SD = 0.18), higher than for those who interacted with fixed answers and delays (M = 0.61, SD = 0.23). For participants exposed to reply suggestion buttons, the follow ratio was on average 0.77 (SD = 0.20), higher than for those who were not exposed to suggestion buttons (M = 0.59, SD = 0.19). The same test revealed no interaction effect (p = 0.325, F = 0.987) between the two variables (see Figure 4).

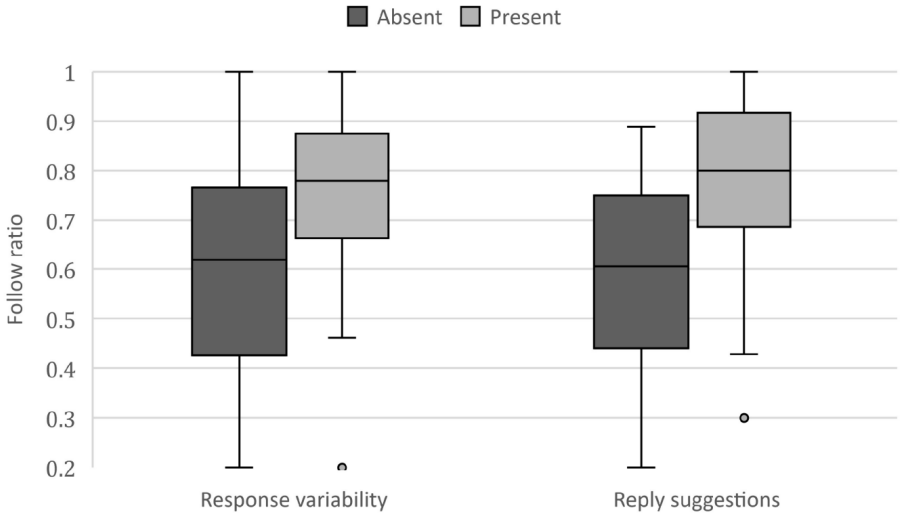

Figure 3: Distribution of the follow ratio across conditions

The ANOVA results on action size revealed a significant effect of response variability (p = 0.019, F = 5.847). Response variability encouraged participants to take actions with greater amounts when following the agent's advice (M = £167.10, SD = £47.78) compared to no variability (M = £117.30, SD = £47.78). The difference across RS/NRS was not significant (p = 0.592, F = 0.290) and there was no interaction effect (p = 0.441, F = 0.601).

A generalised linear model analysis (GZLMA) was applied to the number of messages sent to the chatbot, the number of actions performed, the number of image tags attempted and the number of correct tags, because these variables produced discrete data.

Participants in the RV level sent, on average, fewer messages (M = 55.06, SD = 18.79) than those in the NRV level (M = 64.91, SD = 25.72). The difference is significant (p = 0.015). In contrast, participants in the RS levels sent more messages (M = 71.25, SD = 20.97) than participants in the NRS level (M = 48.72, SD = 19.05). The difference is highly significant (p < 0.001). There was no interaction effect (p = 0.092).

Participants performed fewer actions on portfolios (M = 19.44, SD = 7.148) in the RV level compared to those in the NRV level (M = 26.09, SD = 9.72). The difference is significant (p = 0.001). Reply suggestion buttons had no significant effect on the number of actions performed (p = 0.781) and there was no interaction effect (p = 0.113).

The GZLMA on the number of image tags attempted revealed a significant effect of reply suggestions (p = 0.003), but not of response variability (p = 0.882), and no interaction effect (p = 0.220). Participants with access to reply suggestions made more attempts (M = 101.80, SD = 72.69) than participants with no reply suggestions available (M = 54.81, SD = 53.09).

There was a significant effect of reply suggestions (p = 0.006) on correct image tags, but not of response variability (p = 0.603), and no interaction effect (p = 0.126). Participants with access to reply suggestions guessed more tags correctly (M = 45.19, SD = 34.50) than participants in the NRS level (M = 24.97, SD = 24.98).

No other significant effects on Participant behaviour was found.

4.3 Discussion

The results show that response variability encouraged participants to follow the advice of the chatbot more frequently. The effect of response variability was also significant on the average size of actions following the chatbot's advice, with participants performing actions with greater amounts of money in the RV level. Previous work reveals that agents with variable answers are perceived as more human-like [2, 29, 39] and that the more human-like the chatbot is perceived to be, the more competent it seems to its users [13]. There is clearly much more to anthropomorphism and believability than response variability, but varying utterances and delays appear to increase perceived competence of chatbots nonetheless.

Participants more frequently followed the advice offered by a chatbot with reply suggestion buttons (independently of response variability). Because the suggestions are embedded in the interface of the chatbot, it is possible that this provided a sense that suggestions were generated by the agent itself, which may have demonstrated competence within the social trading environment. Prior research suggests that the perception of competence affects benefit/risk-sharing efforts [23], so it may be that participants in the RS level were more confident that the agent's advice provided a beneficial benefit-risk balance. These findings are in line with earlier work from García et al. [20], who found that simple UI variations do not have to evoke a greater perception of human-likeness to increase the perception of competence of smart systems.

Interestingly, despite following the advice of the agent more frequently, participants in the RS level did not invest or withdraw greater amounts when doing so. Reply suggestion buttons were shown at every stage of the conversation, including moments where participants were asked to determine the size of trading actions. Here, the suggested amounts were based on proportions of the available or invested balance, depending on the action. Given that suggestions were used, on average, 78.44% of the time (SD = 18.09%),

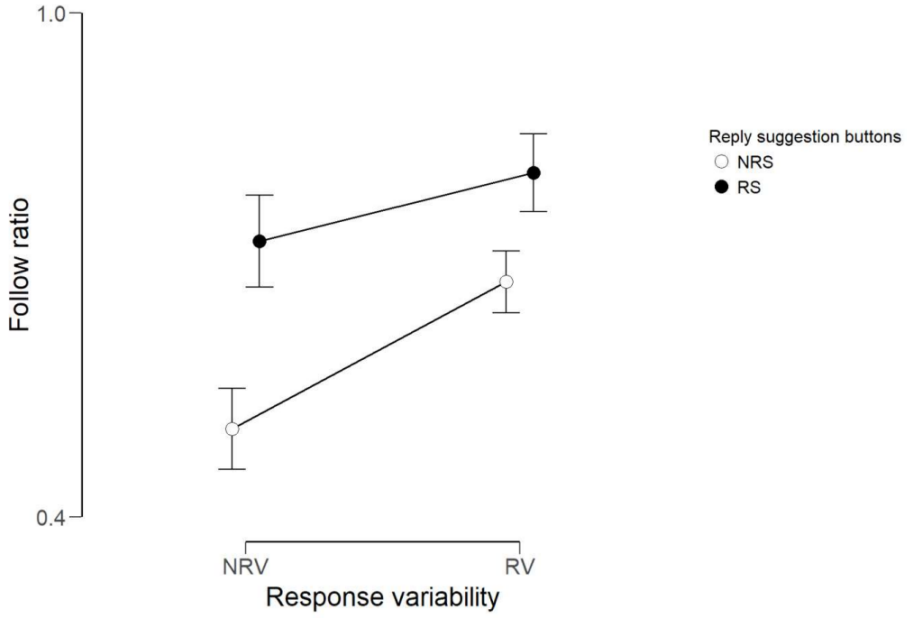

Figure 4: Follow ratio means across conditions

if selected, suggested amounts were similar to the amounts specified by the participants in the NRS condition.

Participants with access to reply suggestions made more attempts at image tagging than those without access to the buttons. The chatbot was kept visible in the image-tagging task view of the platform and the presence of the buttons reduced the time required to shift the focus away from the task; arguably a classic example of increased usability. More generally, it could be argued that reply suggestions allow more seamless integration of chatbot interactions within a multitasking environment.

As expected, reply suggestion buttons also encouraged participants to send more messages as the users were not required to type any text at any stage of the conversation. Participants sent more messages when responses were not varied. Because chatbot utterances in the RV level varied for each message, participants likely required more time to understand what the agent was saying compared to participants who faced the same utterances throughout the experiment. Additionally, messages in the RV level experienced a longer delay, meaning that participants spent a greater total portion of their time waiting for the chatbot to reply.

The average use of reply suggestions for each message (78.44%) is significantly greater than in previous studies performed on Google's email application Inbox designed for mobile devices [27]. There, Smart Replies were responsible for assisting with 10% of email replies. However, our participants interacted with a chatbot, not other humans. Emails are likely to be composed more carefully, especially in a professional context where the accuracy of suggestions is currently insufficient to replace manual typing. Our value is closer to the one found by Gao et. al [19] in their study on suggestion buttons used in casual question-answering conversation (44.2-68.3%).

Overall, these findings suggest that the perception of chatbots is subject to what appears to be a *cognitive bias* and is influenced by relatively simple UI variations. This observation extends recent work that has drawn attention to similar types of cognitive bias in text recognition systems and biosignal sensors [20, 25]. Designers of intelligent conversational agents should carefully consider the implications of novel features, such as response variability and reply suggestion buttons, that may influence the perception of the system and user behaviour significantly. As AI-driven systems are increasingly deployed and adopted in the healthcare sector in clinical imaging, therapy for anxiety and depression, COVID-19 patient management, and more, potential risks of over-reliance might have serious consequences [5, 7, 8, 14]. Indeed, Calisto et al. [6] revealed that clinicians are more likely to expect higher performance from a system that generates easy and understandable recommendations, and that is easy to use.

Our study has several *limitations*. We did not make hypotheses about relationships between participant demographics (e.g. sex, or age) and their behaviour: an opportunity for further research. We also did not collect information about the participants' prior experience with chatbots or investment systems, factors that might influence their perception and beliefs about our system. A post-task questionnaire could have been included to better understand these aspects. Additionally, in contrast to studies where chatbots and prototypes were deployed in real-life scenarios [11, 14, 37], our study was conducted on a simulated social trading platform, which allowed us to compare alternative conditions, but may not have fully captured the complexity and dynamics of real-world investment decisions. Finally, future work could investigate the design aspects of reply suggestions in more detail, such as the order of appearance and content neutrality.

5 Conclusion

This paper reported on a between-subject online study designed to evaluate the effects of response variability and reply suggestion buttons on user behaviour around a chatbot, specifically whether users follow the advice offered by the agent. The study took place on a simulated social trading platform where 64 participants interacted with a chatbot to receive advice on financial investment. The study design allowed us to measure how frequently and to what extent participants followed the advice of the chatbot compared to an alternative source, a news feed, under different between-groups conditions. Through a 2x2 factorial design, we evaluated the effect of response variability and reply suggestion buttons. We employed performance-based financial incentives to make the consequences of following the chatbot's advice tangible to participants.

The results indicate that both response variability and reply suggestion buttons had significant effects on user behaviour around the chatbot. Participants followed the advice of the chatbot more frequently when answers and delays were varied and when suggestion buttons were available.

As chatbots continue to rise in popularity, task-oriented agents are increasingly employed in customer support, counselling, investment and tourism, to name a few. Designers should look at response variability and reply suggestion buttons as ways to increase the efficacy and persuasiveness of chatbots when giving advice to users.

Finally, the findings of this study confirm the importance of investigating how cognitive biases might influence the interaction with complex, intelligent systems. The users' perceptions of chatbots are subject to bias and influenced by relatively simple UI variations. Considering recent work that similarly called attention to bias in relation to smart systems, we hope that this paper will stimulate further work in this area.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council A-IoT (EP/N014243/1) project. The study was approved by the Ethics Committees of UCLIC. We would like to thank all the participants of the study as well as the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. Data URI: https://doi.org/10.5522/04/c.6675386.

References

[1] M. Ammann and Nic Schaub. 2016. Social Interaction and Investing: Evidence from an Online Social Trading Network. https://www.semanticscholar.org/paper/Social-Interaction-and-Investing%3A-Evidence-from-an-Ammann-Schaub/2ac3fe12070ccc1ee6ceec1c2b3e2bfe1f088d40

[2] Jana Appel, Astrid von der Pütten, Nicole C. Krämer, and Jonathan Gratch. 2012. Does Humanity Matter? Analyzing the Importance of Social Cues and Perceived Agency of a Computer System for the Emergence of Social Reactions during Human-Computer Interaction. *Advances in Human-Computer Interaction* 2012 (Aug. 2012), e324694. https://doi.org/10.1155/2012/324694 Publisher: Hindawi.

[3] Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85 (Aug. 2018), 183–189. https://doi.org/10.1016/j.chb.2018.03.051

[4] Warren J. Bilkey. 1953. A Psychological Approach to Consumer Behavior Analysis. *Journal of Marketing* 18, 1 (July 1953), 18–25. https://doi.org/10.1177/002224295301800103 Publisher: SAGE Publications Inc.

[5] Francisco M. Calisto, Alfredo Ferreira, Jacinto C. Nascimento, and Daniel Gonçalves. 2017. Towards Touch-Based Medical Image Diagnosis Annotation. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces (ISS '17)*. Association for Computing Machinery, New York, NY, USA, 390–395. https://doi.org/10.1145/3132272.3134111

[6] Francisco Maria Calisto, Nuno Nunes, and Jacinto C. Nascimento. 2022. Modeling adoption of intelligent agents in medical imaging. *International Journal of Human-Computer Studies* 168 (Dec. 2022), 102922. https://doi.org/10.1016/j.ijhcs.2022.102922

[7] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. 2021. Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies* 150 (June 2021), 102607. https://doi.org/10.1016/j.ijhcs.2021.102607

[8] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. 2022. BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. *Artificial Intelligence in Medicine* 127 (May 2022), 102285. https://doi.org/10.1016/j.artmed.2022.102285

[9] Ana Paula Chaves. 2020. Should my Chatbot be Register-Specific? Designing Appropriate Utterances for Tourism. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3334480.3375033

[10] Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's How You Say It: Identifying Appropriate Register for Chatbot Language Design. In *Proceedings of the 7th International Conference on Human-Agent Interaction (HAI '19)*. Association for Computing Machinery, New York, NY, USA, 102–109. https://doi.org/10.1145/3349537.3351901

[11] Zhifa Chen, Yichen Lu, Mika P. Nieminen, and Andrés Lucero. 2020. Creating a Chatbot for and with Migrants: Chatbot Personality Drives Co-Design Activities. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. Association for Computing Machinery, New York, NY, USA, 219–230. https://doi.org/10.1145/3357236.3395495

[12] Minjee Chung, Eunju Ko, Heerim Joung, and Sang Jin Kim. 2020. Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research* 117 (Sept. 2020), 587–595. https://doi.org/10.1016/j.jbusres.2018.10.004

[13] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems* 92 (March 2019), 539–548. https://doi.org/10.1016/j.future.2018.01.055

[14] Paula Maia de Souza, Isabella da Costa Pires, Vivian Genaro Motti, Helena Medeiros Caseli, Jair Barbosa Neto, Larissa C Martini, and Vânia Paula de Almeida Neris. 2022. Design recommendations for chatbots to support people with depression. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems (IHC '22)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3554364.3559119

[15] Daantje Derks, Agneta Fischer, and Arjan Bos. 2008. The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior* 24 (May 2008), 766–785. https://doi.org/10.1016/j.chb.2007.04.004

[16] Philipp Doering, Sascha Neumann, and Stephan Paul. 2015. A Primer on Social Trading Networks – Institutional Aspects and Empirical Evidence. https://doi.org/10.2139/ssrn.2291421

[17] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *Interactions* 24, 4 (June 2017), 38–42. https://doi.org/10.1145/3085558

[18] Mingkun Gao, Xiaotong Liu, Anbang Xu, and Rama Akkiraju. 2021. Chatbot or Chat-Blocker: Predicting Chatbot Popularity before Deployment. In *Designing Interactive Systems Conference 2021 (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1458–1469. https://doi.org/10.1145/3461778.3462147

[19] Zihan Gao and Jiepu Jiang. 2021. Evaluating Human-AI Hybrid Conversational Systems with Chatbot Message Suggestions. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 534–544. https://doi.org/10.1145/3459637.3482340

[20] Pedro García García, Enrico Costanza, Jhim Verame, Diana Nowacka, and Sarvapali D. Ramchurn. 2021. Seeing (Movement) is Believing: The Effect of Motion on Perception of Automatic Systems Performance. *Human–Computer Interaction* 36, 1 (Jan. 2021), 1–51. https://doi.org/10.1080/07370024.2018.1453815 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07370024.2018.1453815.

[21] Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. 2018. Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. *Research Papers* (Nov. 2018). https://aisel.aisnet.org/ecis2018_rp/113

[22] Eun Go and S. Shyam Sundar. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* 97 (Aug. 2019), 304–316. https://doi.org/10.1016/j.chb.2019.01.020

[23] Byoung-Chun Ha, Yang-Kyu Park, and Sungbin Cho. 2011. Suppliers' affective trust and trust in competency in buyers : Its effect on collaboration and logistics efficiency. *International Journal of Operations & Production Management - INT J OPER PROD MANAGE* 31 (Jan. 2011), 56–77. https://doi.org/10.1108/01443571111098744

[24] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. https://doi.org/10.48550/arXiv.1705.00652 arXiv:1705.00652 [cs].

[25] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On Being Told How We Feel: How Algorithmic Sensor Feedback Influences Emotion Perception. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2 (Sept. 2018), 1–31. https://doi.org/10.1145/3264924

[26] T. M. Holtgraves, S. J. Ross, C. R. Weywadt, and T. L. Han. 2007. Perceiving artificial social agents. *Computers in Human Behavior* 23 (2007), 2163–2174. https://doi.org/10.1016/j.chb.2006.02.017 Place: Netherlands Publisher: Elsevier Science.

[27] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 955–964. https://doi.org/10.1145/2939672.2939801

[28] Dan Kim, Donald Ferrin, and Raghav Rao. 2008. A Trust-Based Consumer Decision-Making Model in Electronic Commerce: The Role of Trust, Perceived Risk, and Their Antecedents. *Decision Support Systems* 44 (Jan. 2008), 544–564. https://doi.org/10.1016/j.dss.2007.07.001

[29] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. Association for Computing Machinery, New York, NY, USA, 555–565. https://doi.org/10.1145/3064663.3064672

[30] Bart Larivière, David Bowen, Tor W. Andreassen, Werner Kunz, Nancy J. Sirianni, Chris Voss, Nancy V. Wünderlich, and Arne De Keyser. 2017. "Service Encounter 2.0": An investigation into the roles of technology, employees and customers. *Journal of Business Research* 79 (Oct. 2017), 238–246. https://doi.org/10.1016/j.jbusres.2017.03.008

[31] Kurt Lewin. 1943. *Forces Behind Food Habits and Methods of Change.* National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK224347/ Publication Title: The Problem of Changing Food Habits: Report of the Committee on Food Habits 1941–1943.

[32] Bingjie Liu and S. Shyam Sundar. 2018. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychology, Behavior and Social Networking* 21, 10 (Oct. 2018), 625–636. https://doi.org/10.1089/cyber.2018.0110

[33] François Mairesse and M. Walker. 2008. Can Conversational Agents Express Big Five Personality Traits through Language ? : Evaluating a Psychologically-Informed Language Generator. https://www.semanticscholar.org/paper/Can-Conversational-Agents-Express-Big-Five-Traits-%3A-Mairesse-Walker/d6a0e683ea321dcfcd52c9be78180079ccaeb424

[34] Inaki Maurtua. 2009. *Wearable Technology in Automotive Industry: from Training to Real Production.* IntechOpen. https://doi.org/10.5772/7742 Publication Title: Human-Computer Interaction.

[35] Wei Pan, Yaniv Altshuler, and Alex (Sandy) Pentland. 2012. Decoding Social Influence and the Wisdom of the Crowd in Financial Trading Network. *MIT Web Domain* (Sept. 2012). https://dspace.mit.edu/handle/1721.1/80764 Accepted: 2013-09-16T20:10:18Z ISBN: 9781467356381 Publisher: Institute of Electrical and Electronics Engineers (IEEE).

[36] J. Paul Peter and Lawrence X. Tarpey. 1975. A Comparative Analysis of Three Consumer Decision Strategies. *Journal of Consumer Research* 2, 1 (1975), 29–37. https://www.jstor.org/stable/2489044 Publisher: Oxford University Press.

[37] Rifat Rahman, Md. Rishadur Rahman, Nafis Irtiza Tripto, Mohammed Eunus Ali, Sajid Hasan Apon, and Rifat Shahriyar. 2021. AdolescentBot: Understanding Opportunities for Chatbots in Combating Adolescent Sexual and Reproductive Health Problems in Bangladesh. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445694

[38] S. Shyam Sundar, Eun Go, Hyang-Sook Kim, and Bo Zhang. 2015. Communicating Art, Virtually! Psychological Effects of Technological Affordances in a Virtual Museum. *International Journal of Human–Computer Interaction* 31, 6 (June 2015), 385–401. https://doi.org/10.1080/10447318.2015.1033912 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2015.1033912.

[39] Jhim Kiel M. Verame, Enrico Costanza, and Sarvapali D. Ramchurn. 2016. The Effect of Displaying System Confidence Information on the Usage of Autonomous Systems for Non-specialist Applications: A Lab Study. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4908–4920. https://doi.org/10.1145/2858036.2858369

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

[41] Mao Xuetao, François Bouchet, and Jean-Paul Sansonnet. 2009. Impact of agent's answers variability on its believability and human-likeness and consequent chatbot improvements. *Proceedings of AISB* (2009), 31–36.