Investigating the Effects of Reply Suggestions on User Trust in Chatbot Applications

Federico Milana
HCI-E MSc Final Project Report 2019
UCL Interaction Centre, University College London
Supervisor: Dr Enrico Costanza

ABSTRACT
User trust is a crucial component to take into consideration in the design of conversational agents as these services increase in popularity and demand. In order to serve a myriad of practical purposes, chatbots are employed in an increasing number of environments to provide a conversational method of interaction. As customer service remains the most popular functionality provided by these systems, the design of automated and personalised interactions has employed a variety of techniques to facilitate customer support. Particularly relevant are assisting writing tools that aim to ease the communication between a person and a natural language processor. In this sense, reply suggestions are integrated to provide an immediate understanding of the chatbot's functionality. However, as these suggestions are deeply contextual to the conversation taking place, it is reasonable to assume some degree of effect on interaction that goes beyond their intended purpose. This study explored the effects of the presence of reply suggestions on trust by analysing the behaviour of users in a situation characterised by risk. To this end, a simulated social trading platform was hosted on a web server whereby participants could interact with a chatbot assistant to receive advice on investment. The platform showcased portfolios whose value changed unpredictably over a fixed time period. In a between-subject design, reply suggestions were displayed to one group of participants and hidden to the other. By analysing the extent to which the advice given was followed or rejected, the study has collected results that suggest that reply suggestions encourage users to trust the chatbot more frequently and in riskier situations.

MSC HCI-E FINAL PROJECT REPORT
Project report submitted in part fulfilment of the requirements for the degree of Master of Science (Human-Computer Interaction with Ergonomics) in the Faculty of Brain Sciences, University College London, 2019.

Author Keywords
chatbot; reply suggestions; interface design; user trust; social trading; online study.

ACM Classification Keywords
H.5.m. *Information interfaces and presentation*: Miscellaneous

MSc Contribution Type
Empirical

1. INTRODUCTION

Assisting writing tools have recently grown in popularity in today's messaging services, ranging from keyboard suggestions to gestural methods of input. Thanks to the implementation of artificial intelligence able to understand the context of a conversation, companies such as Google and Facebook have successfully integrated reply suggestions in their chat services. *Smart Replies* [54], or *Quick Replies* [55], are effectively part of the interface so that, when the user receives a message, they are also presented with a small number of buttons each displaying a contextually relevant answer. Selecting one allows the user to immediately send a message without having to type any text.

Meanwhile, as chatbots continue to advance in terms of natural language processing, they are able to serve an increasing number of practical purposes. These include customer service, virtual assistant interaction and information acquisition to name a few. By offering significant advantages such as instant, 24-hour responses and the lack of dedicated personnel, chatbots have become an attractive method of interaction between users and services.

Studies on online shopping suggest that, more than enjoyment, usefulness and ease of use, trust is crucial in establishing customer loyalty [8], and that social presence has significant effects on shaping purchasing behaviours because of its contribution to user trust [34]. With further evidence revealing chatbots as digital tools able to offer convenient, personal and unique customer assistance [6], it is reasonable to recognise the reason for which user trust in

these systems has been subject to an extensive amount of research. Design aspects studied include font, response time, human-likeness, explanations and professional appearance, while measurements of trust are equally varied [4,5,17,50]. Being a novel feature, reply suggestions have not experienced the same attention as these aspects and therefore present a promising opportunity for a study. This project aims to follow a similar direction by evaluating the effects of reply suggestions on the interaction with a chatbot, specifically on the extent to which the user trusts the system.

While there are numerous ways to measure trust, one approach considers the fact that it is seen as particularly relevant in situations characterised by risk, where the trustor depends on the actions of the trustee [44]. Generating situations involving risk for the user and requiring trust towards the chatbot has therefore informed the choice of integrating aspects of behavioural economics in this study. In this sense, a market environment is arguably perfect to allow situations of trust to emerge in the presence of equity risk.

It follows that the chatbot should then be integrated in an environment that encourages the user to interact with a market by investing in shares whose value can change over time. The study considered social trading platforms such as eToro [57] and ZuluTrade [72] as suitable methods to simplify the integration of a conversational agent as well as the user's understanding of the system. By allowing investors to automatically, simultaneously, and unconditionally copy the investments of other traders by following their portfolios, these networks reduce the amount of relevant knowledge required from users to participate in investment [51]. This aspect is desirable in the practical terms of minimising the possibility that the user is overwhelmed by the necessary amount of information to process before making a decision. For example, given that risk might attract a certain group of investors, whereas other types of investors prefer less risky options [51], a quantified risk index, typical of social trading platforms [57], is preferable compared to resorting to personal calculations, for inexperienced investors especially.

Social trading opens the door to a significant amount of additional aspects to evaluate, such as the role of trustees as producers of signals of trustworthiness, as emphasised by much of the trust literature [30]. However, this study attempts to limit the effect of these by employing a variety of design choices that aim to limit its scope so that social trading is integrated solely as a tool to facilitate the measurement of user trust in the chatbot.

Recommendation agents (RA) are of particular relevance here, and are defined as software agents that elicit the interests or preferences of individual consumers for products and making recommendations accordingly [52]. While the employment of RA's remains confined to e-commerce [29], it is clear that a conversational agent can act as an RA in a social trading environment by offering the user advice on what actions to perform within the market. Faced with the repeated decision to either follow or disregard the agent's advice, it is possible to obtain an indication of the extent to which the user trusts the system, given the ubiquitous presence of equity risk. An exemplar situation would have the chatbot suggesting the user to invest in a particular portfolio, in which case the amount to invest becomes at risk of diminishing. It is the purpose of this study to investigate whether the presence of reply suggestions, such as "Okay, let's do it" and "Never mind", has any significant effect on the behaviour of the user.

In order to determine the behaviour of the user in a social trading environment with a conversational/recommendation agent, a web application was created and hosted on a web server. An online study was run and, thanks to the implementation of a database, it was possible to record participant data aiming to answer the research question.

Following a literature review of pertinent topics, the document describes the features of the simulated social trading platform in separate sections, after which the main findings of the online study are summarised and discussed.

2. LITERATURE REVIEW

2.1 Trust

Trust can be defined as a psychological state in which an agent willingly and securely becomes vulnerable, or depends on, a trustee, having taken into consideration the characteristics of the trustee [24]. As trust characterises the dependent variable of the study, it is critical to understand its various facets in the context of a conversational agent and its application. Literature in this field posits a major distinction between cognitive and affective trust in service relationships. While the first is a customer's confidence or willingness to rely on a service provider's competence and reliability derived from accumulated knowledge [38], the second can be described as the confidence one places in a partner on the basis of feelings generated by the level of care and concern the partner demonstrates [26,27,41].

The large amount of literature concerning human-likeness [5], politeness [46], empathy [52] and emotional intelligence [47] of conversational agents illustrates the multitude of design choices to consider to increase affective trust amongst users. In order to claim that the possible effect of reply suggestions concerns affective trust more than cognitive trust, a critical assumption has to be made. Namely, that reply suggestions are understood by the user

as the chatbot displaying integrity and benevolence in providing guidance throughout the conversation. As these qualities are of significant importance to RA users [3], it is reasonable to infer that, if this assumption is true, the presence of reply suggestions will have a significant effect on the level of affective trust in the chatbot.

On the other hand, it appears that cognitive trust, which is recognised to be characteristically knowledge-driven [26], will demonstrate a strong correlation with the chatbot's performance, as the accumulated knowledge of previous events will allow the user to make predictions on the accuracy and reliability of the system [27]. Since a social trading environment presents randomness and unknown outcomes as necessary components in generating situations of trust, it follows that the accuracy of the chatbot's advice is also randomised. An alternative design would consider setting a fixed accuracy level so that, for example, the agent's advice is accurate exactly half the time. However, despite reducing the effect of chance on user cognitive trust, this approach is counterproductive when considering that cognitive trust presumes a state of incomplete knowledge [26]. In other words, since predictability is an inevitable outcome of fixed chatbot accuracy and performance, it is preferable to accept chatbot performance as an uncontrolled variable than to eliminate risk altogether.

Additional literature points to the difficulty in measuring trust, referring to the inadequacy of research in interpreting behaviour in trust games, such as the prisoner's dilemma, or the differences between guidelines for measuring trust in organisations. The first case presents the issue of identifying trust within complex and context-driven behaviours, where players are just as likely to choose to defect to retaliate than to mistrust the other player [35]. The second case illustrates the multi-dimensional nature of trust comprising of subset measurements of the organisation's perceived competence, integrity, dependability and honesty, to name a few [39].

It is safe to assume that the number of actions to choose from within social trading is larger than in a typical trust game, and that the environment is arguably more complex than that of the prisoner's dilemma. However, the absence of competition between two players reduces the complexity of user behaviour significantly by removing determinants of trust such as perceptions of the other player's moral character, individual misfortunes, economic status, cultural and religious roots which lead to different philosophical attitudes toward interactions within trust games [9,1].

The multi-dimensional nature of trust remains an important issue to take into account. However, this characteristic is particularly relevant for cases in which a measurement of trust is to be devised. In the case of this study, quantifying trust by the amount of times the agent's investment advice is followed appears suitable and clear to a satisfying extent.

2.2 Chatbots

The design of chatbots has been subject to an extensive amount of research, mostly covering the effects of interface design and communication techniques on the perception of believability and humanness. Despite not always concerning user trust specifically, most of the literature is still useful in informing the development process of this study. An example is research relating to effects of typefaces on the perception of humanness in chatbots, which indicates a predisposed tendency to perceive conversational agents as such regardless of visual aspects [5]. It follows that, if the chatbot developed in this study aims to display human characteristics, it should do so through its language rather than visual appearance.

Research reveals that, despite featuring shorter messages than in human-human online conversations [23], human-chatbot conversations can still display emotional traits to a comparable extent, with over 40% of customer service requests to chatbots being of this kind [53]. Supporting this is further evidence suggesting that computer-mediated communication is able to communicate emotion as well as or better than face-to-face communication [10]. Nevertheless, as the purpose of the social trading agent is practical in terms of offering advice on which actions to perform, one might argue that handling conversations of emotional nature is of secondary importance. However, research in this field also suggests that expressions of sympathy and empathy are favored over unemotional provision of advice [32]. Therefore, the system should consider the user's expectations to engage in conversation of emotional nature in order to appear trustworthy. While the focus of the study is on the effects of the presence of reply suggestions, it is clear that the agent is required to demonstrate some degree of trustworthiness regardless of the condition to produce meaningful results.

A literature review of chatbots in the context of user trust cannot underestimate the importance of explainable artificial intelligence (XAI). XAI aims to achieve complete trustworthiness and an evaluation of the ethical and moral standards of a machine by giving detailed explanations of AI decisions [13]. Such explanations should provide insight into the rationale the AI uses to draw a conclusion [50]. Given that explanations of one's decisions are often a prerequisite for establishing a trust relationship between people [43], it follows that a social trading RA should provide motivations behind every piece of advice given. However, it is important to put this into the perspective of the study. The explanation of the advice given is likely to surpass the knowledge of the user, who is not expected to be a trading expert. Moreover, the integration of XAI defies

the presumption of a state of incomplete knowledge in order to create situations of cognitive trust. In other words, as explanations significantly increase the trustworthiness of the agent, they excessively reduce the amount of trust required from the user. This reflection informs a clear decision of disregarding elements of XAI, which offer a promising direction for a different study nonetheless.

On the other hand, literature on response delays of conversational agents indicate that delays not only increase users' perception of humanness and social presence, but also lead to greater satisfaction with the overall chatbot interaction [18]. However, the presence of delays is not the only design aspect to consider here, as research also suggests that users expect delay times to vary depending on the content of both the request and response [36]. Therefore, in order to replicate a successful implementation of an agent that is believable and meets user expectations, the system requires response delays to exist and vary in length.

2.3 Reply Suggestions

Literature on reply suggestions is lacking, but research made on Google Inbox's *Smart Reply* feature is relevant. 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 [28]. In contrast to an implementation within a conversational agent, where reply suggestions can be generated according to the messages sent by the chatbot, *Smart Replies* for emails require an acceptable degree of response prediction accuracy. The generation of reply suggestions in chatbots require no neural natural language understanding models, which removes a significant challenge in development [22]. For this reason, while the *Smart Reply* system was responsible for assisting with 10% of email replies for *Inbox* on mobile in 2016 [28], the study assumes that this value is much greater in conversations with an agent. Regardless, it would be useful to adopt the same measure of success by recording the amount of times reply suggestions are used.

2.4 Social Trading

While literature on behavioural economics is vast and exhaustive, a relevant review is limited to social trading platforms and analyses available examples such as eToro and ZuluTrade.

Social trading platforms provide access to an innovative type of delegated portfolio management by allowing investors to copy investment strategies of other traders [11]. The mechanisms that emerge from the "copy-trading" technique suggest that, while social trades outperform individual trades, the social reputation of the top traders in eToro is not completely determined by their performance. One study has attributed the absence of correlation between rankings and performance to information regarding the number of followers, or copiers, (see Figure 1) as users are heavily affected by this number when making mirroring decisions even when trading performance is completely disclosed [40].

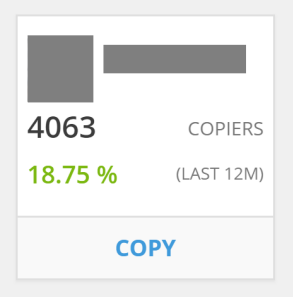

Figure 1. A portfolio thumbnail displayed on eToro (profile picture and username hidden)

This study should take into account the significance of social effects, which can override rational assumptions. In this sense, it would be 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 complex. 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 concerns an increased asymmetry in information held by users and traders. While social trading platforms aim to reduce information asymmetry by publishing a full single transaction history and standardized real-time track records for each portfolio [11], this is likely to detriment an analysis of data produced by the 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 traders' communication impacts investment decisions of followers, despite comments not containing informational value [2]. In order to simplify the interpretation of actions from the user, the study considers this to be 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 [58,72], which are advertised as primary features, forbidding access to this information would excessively affect the essence of social trading.

A final note is to be made regarding conversational agents within the context of trading. Interestingly, it appears that a similar agent mentioned by this study has been developed by the Bombay Stock Exchange in partnership with

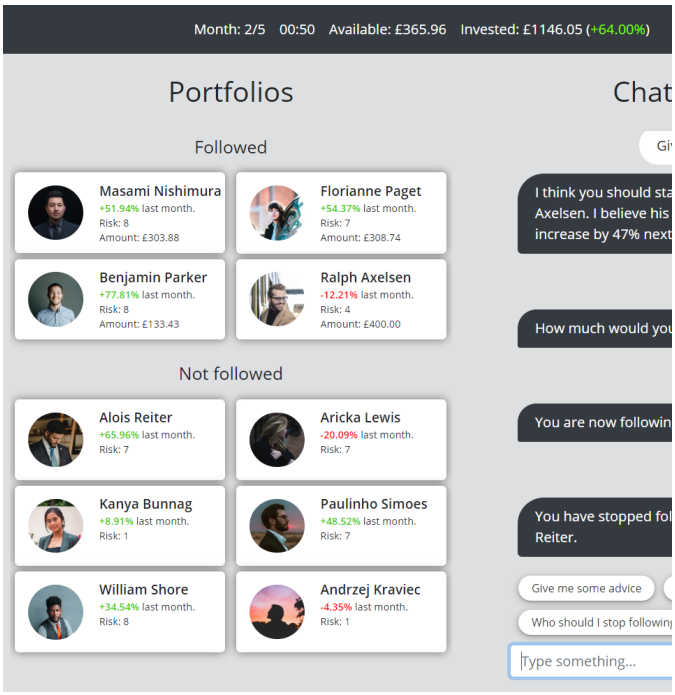

Figure 2. The main view of the web application

Microsoft and ShepHertz in 2018 [38]. The chatbot, called "Ask Motabhai", is able to provide on-demand data and information from the stock exchange's website. However, aiming to serve the audience with quick information without having to go through any other channel, its functionality is limited compared to that of an RA, which is also capable of offering advice.

3. SIMULATED SOCIAL TRADING PLATFORM

The concepts of trust, chatbots, reply suggestions and social trading explored in the literature review informed the design of a simulated social trading platform (see Figure 2) whereby the integration of a conversational agent could produce data aiming to answer the research question of the study. In order to fully understand the nature of the platform, its major design aspects are described in the following sections.

3.1 Portfolios

The centerpiece of any social trading platform, 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, over time, its value can change, depending on the performance of its respective trader. Portfolios required decisions to be made in terms of design choices and implementation within the platform, which are described below.

Simulating a Fluctuating Market

It is clear that some sort of mechanism has to be employed to determine the change in value for each portfolio in the platform. This aspect is of significant importance in a simulation when considering that fluctuations need to adhere to an element of randomness to maintain a degree of unpredictability of the environment.

It was decided that the platform would aim to achieve this by leveraging on a key feature of real social trading platforms: the risk score. In order to facilitate the assessment of risk associated with trading decisions, platforms calculate a score between 1 and 10, where 1 is the lowest possible risk and 10 is the highest possible risk [59]. It follows that, while the calculation of this score is based on the trader's past performance, the risk score provides an indication of the magnitude of possible future gain. This rules out an alternative approach that would allow the value of portfolios to change solely based on randomness. (It is important to note here that change in value, or gain, is intended in percentages and not in net difference.)

In order to achieve random fluctuation while taking risk into account, the simulation calculates the change in value for each portfolio by generating a random normal/gaussian variable with a mean of 0.00 and standard deviation equal to a multiple of the risk score. This ensures that the probability of the value of a portfolio with a low risk score to change dramatically is lower than the probability of the value of a portfolio with a high risk score. At the same

time, the "bell shape" of a gaussian distribution ensures that change tends towards the mean value of 0.00 so that positive or negative dramatic change is less likely to occur than smaller fluctuations. In order to maintain a degree of realism, limits to change were set to -99.99 and +99.99.

While the value of portfolios in real social trading platforms experiences continuous change over time, the time constraint of an eventual study run on a simulated platform informed the decision of allowing change to occur discretely over a set period of "months". In this implementation, values of portfolios change after the passing of each month. An alternative approach would have seen values fluctuating continuously, but the time constraint would have required changes to be significantly larger than those of real portfolios. Not only would this design decrease perceived realism of the platform, but it would also require users to perform actions as fast as possible in order to avoid immediate losses. Instead, the decision to implement 5 4-minute months allow the user enough time to engage with the entirety of the system, including the social trading platform, the conversational agent and additional features of the web application that will be described later.

Providing Methods of Interaction

The nature of a social trading environment implies the availability of a disposable balance to invest in portfolios. To simulate an appropriate amount based on typical behaviour on real platforms, users are given a sum of £1000.00 as the initial available virtual balance. Considering the implementation of real platforms, where users can decide the amount to invest in each portfolio, the actions of *following* and *unfollowing* appear integral to a trading environment and are replicated by the simulation. Just like in real platforms, unfollowing a portfolio does not foresee any penalties or fees and simply allows the user to withdraw the total amount invested in a portfolio.

However, one noteworthy difference in terms of interaction with portfolios concerns the possibility to withdraw only a portion of the invested amount. While this action is not allowed on real platforms, where the user can simply copy or stop copying portfolios, the decision was informed by the inclusion of a time constraint. Without the ability to easily move invested amounts from one portfolio to another, users would waste unnecessary time in a sequence of follow/unfollow actions. For the same reason, users are allowed to add amounts to portfolios that are being followed already. In summary, the actions that the user can perform in relation to portfolios are:

- *Follow*: Start copying the investment of a trader by investing a fixed amount from the available balance
- *Unfollow*: Stop copying the investment of a trader by withdrawing the entirety of the invested amount
- *Add*: Adds a fixed amount to a portfolio that is already being followed
- *Withdraw*: Withdraws a fixed amount from a followed portfolio

Designing Towards Information Asymmetry

Information asymmetry between users and (virtual) traders is desirable when creating situations where trust between the user and the agent is required, as discussed beforehand. Following the decision to simplify the decision-making process of users when participating in social trading, the type and amount of information displayed about each portfolio was carefully selected (see Figure 3).

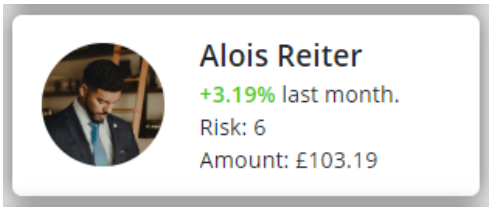

Figure 3. A portfolio displayed in the web application

While communication from traders was discarded, information about risk was kept accessible to users. The reason for this is that, as revealed by the review of literature related to social trading platforms, the effect of comments from traders has a significant effect on user behaviour that is difficult to interpret, as its correlation with informational value is unclear [2]. On the other hand, risk is critical in trading environment, attracting or repelling certain groups of investors [51] and therefore represents a crucial piece of information about traders.

On the other hand, an indication of the portfolio's gain achieved in the previous month is included because of necessity. As the user invests a fixed amount in a trader, this information is accessible regardless of an indication due to the change in the invested amount. The absence of a gain indicator would unnecessarily burden the user with the task of calculating the value personally.

3.2 Chatbot

The chatbot developed by this study is fully embedded in the user interface of the platform, allowing a conversational method of interaction with the social trading environment (see Figure 4).

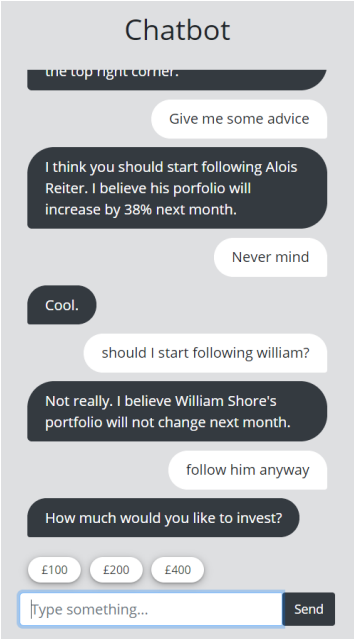

Figure 4. A typical conversation with the chatbot

Assigning a Role to the Agent

While the role of a recommendation agent is clear when the requirement to offer trading advice is considered, a decision was to be made regarding whether the chatbot should also perform actions on behalf of the user. For instance, should the interface of the platform allow the user to follow a portfolio directly by clicking on a "Follow" button in a similar manner as in real platforms? In this case, the chatbot would become a full-fledged recommendation agent, with its functionality limited to offering trading advice. However, the study aims to evaluate the effects of the presence of reply suggestions on user trust. For this reason, the design of the platform should maximise the amount of interactions with the environment where reply suggestions appear, especially those that elicit the most trust, such as deciding the amount to invest in a portfolio. Accordingly, the role of the chatbot was not limited to that of a recommendation agent. Rather, the design of the platform presents the conversation with the chatbot as the only way to interact with the trading environment.

Degree of Perceived Humanness

As revealed by a literature review relating to typical conversations with chatbots, users have expectations regarding the agent's ability to display an appropriate understanding of emotions [53]. Since the study is interested in the interaction between the user and the chatbot, it is crucial to avoid situations where the presence of the agent is disregarded because of a flawed implementation.

The most apparent design choice in this sense was to integrate an understanding of "chitchat", where the user can greet the agent and engage in conversation that is inconsequential in the context of the platform. As demonstrated by the records of the participants' messages exchanged with the chatbot, a significant portion of conversations begin with "chitchat".

The second aspect considered was the integration of dynamic response delay. In order to perform actions on behalf of the user, such as withdrawing from a portfolio, the chatbot is required to interact with the platform by performing a sequence of operations, including natural language processing and database calls. Expectedly, these operations take some time (typically a total of 1-3 seconds), which coincides with an appropriate response delay. However, because not every message sent by the agent requires a sequence of operations, such as replying to a simple "Hello", a delay of 500ms was set in these cases to avoid immediate replies from the chatbot, which appear artificial to users [36].

A typing indicator (see Figure 5) is displayed in the time frame between a user's message and the chatbot's reply for two reasons. Because response delay depends on the time taken by natural language processing and database calls, its length is dynamic and unpredictable to the user. Providing feedback is therefore essential in achieving visibility of system status, a critical usability heuristic [25]. Additionally, the presence of the indicator aims to replicate the user interface of chat applications used between humans so that the system implies that the agent is typing a message rather than performing background operations.

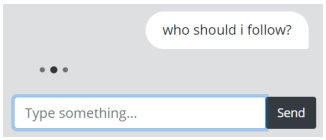

Figure 5. Typing indicator for dynamic response delay

Offering Advice

In order to measure user trust, the study requires the chatbot to provide the user trading advice that is based on the agent's predictions on future portfolio performance. As discussed earlier, the accuracy of predictions needs to be considered carefully, as the performance of the chatbot has significant effects on the user's cognitive trust [27].

The approach taken allows two gaussian random variables to be generated by the system for each portfolio each month. While the agent has access to one variable, the other is only accessible to the newsfeed, which will be described later. At the end of each month, one of the two variables is randomly selected to be the actual change in value for each

portfolio. This ensures that the advice offered by the chatbot is only accurate about half the time, meaning that the user should have no incentive to trust or mistrust advice based on performance. Of course, this ignores situations where advice happens to be always accurate or inaccurate in a given month, which can affect the level of user trust in the agent. However, controlling the accuracy of predictions would hinder elements of randomness and unpredictability, which are necessary components of trading.

Furthermore, it was decided to maintain a degree of simplicity in the kind of advice that the chatbot should offer. By having access to one of the gain values for each portfolio, the agent is able to predict which trader will experience the greatest positive or negative change. It follows that advice is to be given in the context of these values so that, if the user would like to know which portfolio to follow, the advice given relates to the trader with the highest predicted growth that is not being followed already. The decision to limit advice to following and unfollowing portfolios was informed by an eventual interpretation of the actions performed by the user. For instance, if the user decided to unfollow a trader in the third month, it would be possible to look at the change value of the respective portfolio accessed by the agent in the third month to determine whether advice was followed.

Possible advice given by the agent consist of:

- *Following advice*: Suggests the user to start following the portfolio with the highest predicted gain
- *Unfollowing advice*: Suggests the user to stop following the portfolio with the highest predicted loss
- *General advice*: Suggests the user to start following an unfollowed portfolio or stop following a followed portfolio, whichever has the highest change value (see Figure 6)
- *Individual advice*: Suggests whether the user should start or stop following a given portfolio

In order to maximise the number of times the user is offered advice, the design of the chatbot takes into account the possibility that the user never asks for advice. This case does not generate any situations of trust and is therefore undesirable in the scope of the study. To address this, periodic general advice is given at 45 second intervals. Since periodic advice is likely to interfere in a conversation, the system resets the counter after each message is sent by the user.

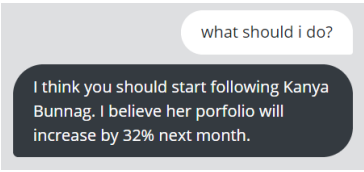

Figure 6. General advice offered by the chatbot

Reply Suggestions

Compared to the implementation of reply suggestions in email applications which requires natural language processing based on the content of the email received, the integration of suggestions in the chatbot presented a different challenge. Given that these should follow the flow of the conversation taking place, it was important to acknowledge the importance of whether the system should only display suggestions that matched the agent's advice, when given. For example, if the chatbot informs the user that they should start following a particular portfolio, should the interface display one reply suggestion containing the text "Do it", or should an additional, contrary suggestion "Never mind" also be displayed? To answer this question, the study assumes that the motivation for using reply suggestions lies in the conveniency of not having to type any text. It is clear, then, that if this conveniency is only available to decisions where the agent's advice is followed, the users' actions do not necessarily reflect a decision of trust in the chatbot. Rather, the decision to follow the advice given would also derive from the absence of a reply suggestion that disregards the advice. For this reason, suggestions displayed in the interface maintain a degree of neutrality by including all kinds of possible replies that are relevant to the current conversation flow (see Figure 7).

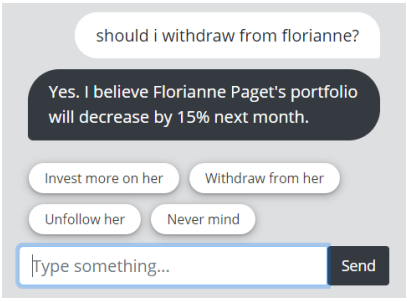

Figure 7. Reply suggestions displayed to the user

3.3 Newsfeed

The decision to integrate a newsfeed (see Figure 8) in the social trading platform was derived from previous design choices relating to information asymmetry between users and virtual traders. As the inclusion of communication from traders was discarded, the need to access information outside the chatbot remained to be met. While real platforms disclose exhaustive data on previous performance for each trader, its interpretation relies excessively on the user's knowledge on the intricacy of trading. Compared to generating predictions based on past performance data, reading posts from a newsfeed appears as a simpler and more intuitive method to inform a trading decision.

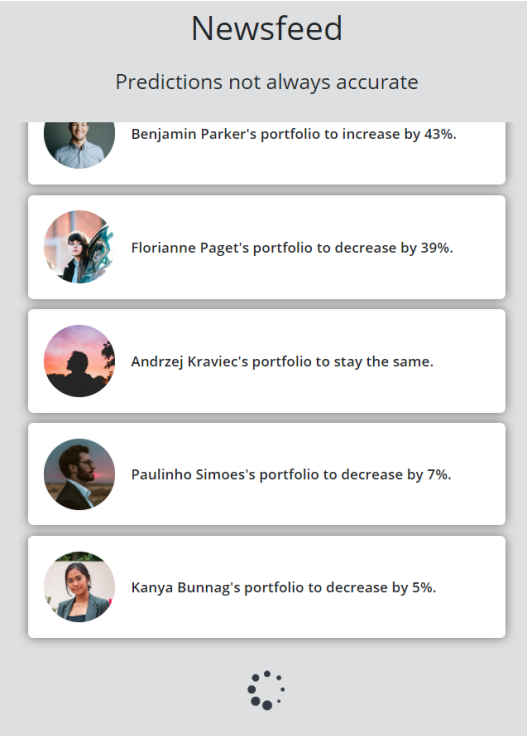

Figure 8. The Newsfeed section of the main view

In the final implementation of the newsfeed, each news post contains a prediction of a portfolio's future gain. It is made clear to the user that posts are have a different source of information than the chatbot, and that predictions between the two might not coincide. In fact, of particular interest to the study are the cases in which the chatbot's advice and a prediction of the newsfeed are contradictory, where one predicts a positive change and the other suggests a negative change in value. It is in this case that the user can indicate their level of trust in the chatbot by deciding to either trust the agent or the news.

One might argue here that the newsfeed and the chatbot should always provide opposite predictions to generate as many situations of trust as possible. However, such an approach would decrease the overall realism of the simulation, as well as possibly increasing the likelihood of the user understanding the mechanism employed to generate portfolio gain. Chatbot and newsfeed predictions are therefore not always contradictory. In a similar manner to the chatbot, the newsfeed generates predictions for each portfolio based on one of the two change values, which has a probability of 50% to be selected as the actual value by the end of the month.

It is crucial to note that this approach entails a difference in the degree of portfolio gain between the chatbot's advice and the newsfeed's predictions. In other words, the chatbot might suggest that the value of a particular portfolio is likely to increase by 25%, while the newsfeed might predict a change of -3%. It is obvious that the decision to follow the advice of the agent in this case does not require a significant amount of trust from the user. To consider this, an appropriate analysis of trust will have to account for the difference in predicted change values as well as other important factors, such as the amount invested or withdrawn.

A final note on the newsfeed concerns the design choice of displaying posts gradually at a random order throughout the month. Not only does this intend to mimic the behaviour of a real stream of news, but allows for the integration of a secondary task within the platform, which is explained in the following section.

3.4 Secondary Task

Given the presence of time constraint experienced by the user in the interaction with the platform, the integration of loading times for the newsfeed can be of considerable importance.

While the system guarantees that all of the posts are generated by the end of each month so that users can view predictions on every portfolio, it can be argued that the act itself of waiting for the newsfeed to generate a post before deciding to invest in a trader is an indication of mistrust towards the chatbot. In contrast, following the advice offered by the agent before viewing the prediction of the newsfeed can be interpreted as a sign of trust. However, the potential of an additional measure of trust is nullified by the fact that the entirety of the information displayed by the newsfeed is accessible by the end of each month. For this reason, it was decided to include a secondary task that would draw away the attention of the user towards the newsfeed.

To achieve this, the secondary task needs to meet several requirements. Firstly, the task needs to provide data that is measurable when determining the amount of effort invested in order to deduce the extent to which the user's attention was drawn away from the newsfeed. Secondly, it needs to offer some kind of reward which is proportional to the effort it demands. Ideally, this is comparable to the amount the user would gain by waiting for the entirety of the newsfeed to generate in the main page. This leads to an additional requirement, which concerns the fact that news should be hidden from the interface in the time the user spends performing the task. Finally, the user should be able to decide to switch between investing and engaging with the task at any given moment during the simulation. Allowing this would ensure the freedom of the user to allocate time according to their judgement.

Image-Tagging

To meet the numerous requirements of the secondary task, an Image-Tagging activity was chosen and integrated into the platform (see Figure 9). The task can be performed on a separate page, or tab, which hides the entirety of the main view excluding the navigation bar that still provides critical information, such as the time left in the current month, available and invested balance.

In this page, the user is presented with an image and is required to enter three keywords, or tags, that describe it. When three tags are found, the user collects a reward and a new image is displayed. The platform handles situations where the user is stuck by allowing them to proceed to the next image even if the tags are not found.

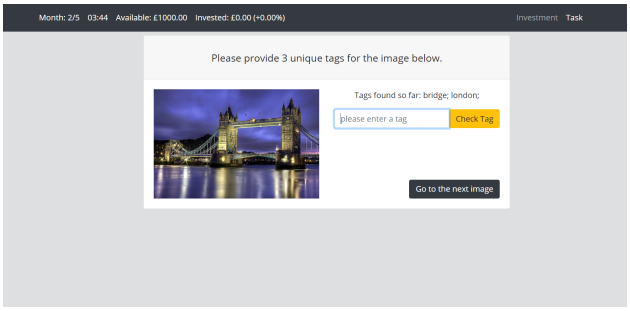

Figure 9. Image-Tagging view of the web application

Calculating Rewards

The amount to reward the user for each image tagged successfully was derived by taking into account the expected time required by each image as well as the expected virtual profit to be made in the investment page of the platform.

In a preliminary study run with a sample (n = 5) recruited from personal connections, the maximum number of images tagged in 20 minutes was found to be 40, indicating a maximum rate of about an image tagged every 30 seconds. On the other hand, the maximum total virtual balance when only engaging with the investment task was found to be £1838.46. By accounting for a comparability between the value of time spent investing and that of time spent tagging, it was possible to calculate £20 as an appropriate virtual reward for each image tagged.

Displaying Notifications

As mentioned beforehand, the newsfeed section is hidden to the user in the page displaying the secondary task. This is intentional, since the purpose of the Image-Tagging activity is to draw the user's attention away from the news. Inevitably, this also reduces the amount of interaction between the user and the chatbot, which is only able to offer trading advice in the main view of the platform.

This issue was addressed by the implementation of periodic notifications that contain trading advice from the agent that appear in the Image-Tagging page while the user is performing the secondary task (see Figure 10). This way, the newsfeed section remains hidden, whereas the interaction with the chatbot can still take place.

Notifications are designed in such a way to allow the user to reply directly without having to switch to the main view. Without this functionality, the user would have to switch back to the investment page to follow the agent's advice, where it would be possible to access information displayed by the newsfeed. Moreover, because this action takes time away from the secondary task, it should only occur in situations where the user does not trust the chatbot. In this case, the users prefers taking time away from a profitable task to check the newsfeed rather than trusting the agent.

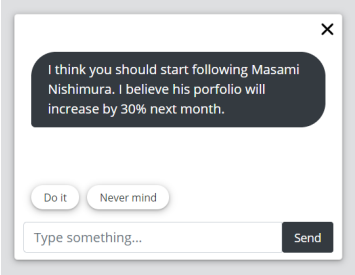

Figure 10. Notification in the Image-Tagging page

4 ONLINE STUDY

4.1 Implementation of the Web Application

The implementation of the web application was achieved with software developed in a *REST* architecture style (Representational State Transfer) [70]. *Python* (v3.6.1)[68] was used to develop the back-end so that *Django* (v2.0.7)[56] could be employed as a web framework. While *Django* makes use of an *SQLite* [71] database, *MySQL* [66] was used once the application was deployed to the UCL server. For the front-end, *JavaScript* [64] was used in combination with *jQuery* (v3.4.1)[65] and *AJAX* requests to handle the responsiveness of *HTML* documents styled with *Bootstrap* (v.4.3.1)[60].

An additional variety of libraries were integrated in the software to facilitate development. The most notable of these is *Rasa* (v1.1.3)[69], an open source machine learning framework to automate text-based conversations. Responsible for natural language processing and dialogue management, *Rasa* was crucial in the development of a functional and contextual chatbot to be used by the study. By providing and classifying examples of possible utterances in the *natural processing unit* (NLU) and of

dialogues as *stories*, it was possible to create agent able to process a conversation with the user in the context of a trading platform.

The final version of the software follows a modular structure containing a module for each component of the platform, as envisioned by *Django* to encourage reusability. These include a *chatbot* module for the functionality presented in the main view of the application, a *chatbotproxy* module to connect *Rasa* with the server, an *imagetagging* module taken from a previous experimental study that was kindly made available for reuse by Dr Enrico Costanza, an *investment_bot* module to assist deployment to the UCL server and a *rasachat* module containing configurations and training models for the chatbot.

Development Process

The development process of the web application applied an incremental build model as a software development method. Phases of development were therefore sequential and repeated until all of the requirements were met. For each requirement, the back-end implementation followed the design of front-end elements after which both aspects were tested.

Development started with the implementation of the social trading logic. Once the interface elements of the portfolios and newsfeed were also integrated and tested, the Image-Tagging page was included by reusing the *imagetagging* module. Subsequent to small adaptations of this module to the structure of the platform was the development of the chatbot and its integration with social trading logic. In order to test the NLU of the chatbot, a pilot study of a small sample (n = 5) recruited from personal connections. The participants' conversations with the chatbot were recorded on the database and were useful to refine the accuracy of the NLU by revealing instances where the user's utterances were not interpreted correctly by the chatbot. After adjusting the unit accordingly, the entire application was hosted on the UCL server by Dr Enrico Costanza.

4.2 Research Questions and Hypotheses

Following the Advice of the Chatbot

Will the presence of reply suggestions in the interface of the chatbot increase the number of times participants follow the agent's trading advice compared to the predictions of the newsfeed?

- *H1*: Participants with reply suggestions will follow the advice of the chatbot more frequently than participants with no access to reply suggestions.

Social Trading Behaviour

Will the presence of reply suggestions affect the amount invested in portfolios in each trading action performed by participants?

- *H2*: Participants with reply suggestions will invest in or withdraw from portfolios greater amounts at a time than participants with no access to reply suggestions.

Degree of Trust Towards the Chatbot

Will the presence of reply suggestions affect the total trust of participants towards the chatbot?

- *H3*: Participants with reply suggestions will display a greater trust towards the chatbot than participants with no access to reply suggestions.

(The measure of trust devised by the study is explained in the *Statistical Analysis Section*)

Secondary Task Management

Will the presence of reply suggestions in the interface of the chatbot increase the effort invested in the Image-Tagging task by participants?

- *H4*: Participants with reply suggestions will invest more effort in the Image-Tagging task than participants with no access to reply suggestions.

4.3 Method

Participants

The main online study recruited a total of 20 participants (7 female, M_{Age} = 25.10, SD = 5.21, range: 18-35) from *Prolific* [67], a crowdsourcing platform that provides a participant pool for online studies. While English was the first language of 10 participants, 15 participants claimed to be fluent. The most common countries of residence were United States (5) and United Kingdom (4). The study recorded the Prolific score of participants (M = 98.75, SD = 3.11, range = 87-100). Of the 20 participants, 6 worked full-time, 6 were unemployed, 3 worked part-time and 2 were not in paid work. Every participant received a minimum compensation of £5 plus an extra £0-£10 depending on their performance in the study. With an average of about 25 minutes to complete the study, participants were paid according to the UK National Minimum Wage as of 2019 [61] excluding bonus rewards, which consisted of £6.57 on average. The only criteria set for participant recruitment was minimum age, already set by Prolific to 18.

Design

The online study used a between-subject design where participants were randomly assigned to either a reply-suggestion condition (RS) and a no-reply-suggestion

condition (NRS). The assignment of the condition to each participant was controlled by the server database which ensured an equal distribution of participants across the conditions.

In the RS condition, participants were shown reply suggestions in the interface of the chatbot, while in the NRS these were hidden. Likewise, reply suggestions in the notification were displayed in the RS condition but hidden in the NRS condition. The social trading logic of the platform, the Image-Tagging task, its reward, the functionality of the chatbot and other aspects of the simulated social trading platform remain unchanged in both conditions.

Dependent Variables

Process variables:

- Number of times a chatbot's advice is followed
- Number of times a newsfeed prediction is followed
- Amount invested when following the chatbot's advice
- Number of tag attempts in the Image-Tagging task

Outcome variables:

- Amount earned/lost through investment
- Amount earned through Image-Tagging
- Final total balance

Procedure

The online study was run on Prolific, which allowed participants to access the web application remotely on their own machines. Due to the method of assigning conditions to participants, both conditions were run simultaneously over a period in which the study was hosted on the crowdsourcing platform. This was done during the evening of one working day.

Each participant was initially presented with a sequence of web pages. These included a *Welcome* page displaying basic information on the content of the following introductory pages: the *Participant Information* and *Informed Consent Form* pages. In the former, a summary relating to the study was provided. This consisted of the purpose of the study, its expected duration, the tasks that participants were expected to perform, the bonus rewards as well as information regarding data usage, conditions for withdrawal from the study and possible ways to express concerns or complaints. The *Informed Consent Form* page provided participants with the contact details of the investigators, an investigator statement confirming that the purpose and context of the study was explained and a participant statement to agree to in order to consent to the terms of the study.

Once consent is given, each participant begins the study after being redirected to the main view of the web application, where the 4-minute timer for the first month is started. The view is split into three sections so that the left column displays 10 portfolios (split into *Followed* and *Not Followed* subsections), the central column contains the interface of the chatbot and the right column shows the newsfeed. Initially, all 10 portfolios are placed under the *Not Followed* subsection of the page. The chatbot greets the participant and offers a brief summary of the messages they can type, while a loading indicator is shown in the newsfeed as the first post is being generated.

The navigation bar at the top displays information on the left about the current month, the time left until the end of the month, the available balance and the invested balance. On the right of the bar, the buttons *Investment* and *Task* are displayed. The Image-Tagging view presents each participant with an image and a textfield where they can enter tags.

As the first timer reaches 0:00, the participant is alerted that a new month has started and that the value of portfolios has changed. At the end of the 5th month, each participant is redirected to a results page displaying a summary of their performance over a period of 5 months, as well as the final total balance.

To complete the study, the participant visits the final page containing a brief questionnaire consisting of 2 Likert-scale and 3 open questions, none of which were optional. The first two questions relate to how much the participant trusted the chatbot and the newsfeed at the by the of the study, while the open questions are based on the overall experience interacting with the chatbot and the platform. After completing the study, participants are redirected to Prolific.

Statistical Analysis Methods

In order to determine whether there is a statistically significant difference between the means of the two unrelated groups, Independent-Samples Student T-tests require some assumptions to be met. Firstly, that the scale of measurement is of a continuous or ordinal nature. This is indeed the case for counts and ratios of the times the participants follow the advice of the chatbot vs the predictions of the newsfeed, as well as the sums of the attempted tags for the secondary task.

Another assumption is that data is collected from a representative portion of the total population. Because the sample of participants was randomly selected with no real criteria, the study assumes this to be the case. The required normality of data was guaranteed by running the *Shapiro-Wilk* test indicating $p>.05$ on all process variables

for both conditions except for tag attempts for participants in the NRP condition. Since the *Shapiro-Wilk* test suggested a deviation from normality for this distribution ($p < .001$), it was decided to use the *Mann-Whitney U* test to compare tag attempts as a nonparametric alternative to the Student's T-test that does not require normality of data.

The assumption related to a homogeneity of variance was tested using *Levene's* test which also yielded non-significant p values for all dependent variables, suggesting that the assumption of equal variance cannot be rejected.

Finally, the results of the Likert-scale questions are analysed using independent-samples student T-tests given that the data collected meets the above mentioned assumptions, including normality of data and homogeneity of variance.

A note needs to be made regarding how data was filtered during the analysis. Because the study was interested in situations where the chatbot's advice and the predictions of the newsfeed were contradictory (one positive and one negative change), trading actions where this was not the case were discarded. Therefore, the results that followed were based solely on cases where the participant had to decide whether to follow the advice of the chatbot or the prediction of the newsfeed.

To test *H3* the study devised a method to take into consideration every aspect of each trading action, namely the difference between the predicted changes, the degree of changes and the amount invested or withdrawn. Following is an explanation of how this was achieved.

Each trading action was classified under *follow*, *unfollow, add* and *withdraw*. According to the predicted (contradictory) changes of the chatbot and the newsfeed, it was possible to determine which advice was followed for each action. A *chatbot trust ratio* for each participant was therefore determined by dividing the number of times the chatbot's advice was followed by the total number of actions where advice was contradictory.

By considering the difference between the predictions of the chatbot and of the newsfeed, it was possible to reveal how much weight each participant had to give to each prediction to balance them exactly:

$$c \cdot p_{chatbot} + n \cdot p_{newsfeed} = 0$$

where c = weight given to the chatbot's advice, n = weight given to the newsfeed's advice and p = predicted relative change of value of the portfolio.

The minimum amount of trust required to follow the chatbot's advice can therefore be quantified as:

$$\frac{c}{n} = -\frac{p_{newsfeed}}{p_{chatbot}}$$

$\frac{c}{n}$ or $\frac{n}{c}$ were calculated for each trading action depending on whether the participant followed the prediction of the chatbot or of the newsfeed. While $\frac{n}{c}$ was kept as a negative value, $\frac{c}{n}$ was always considered as a positive value so that it was possible to calculate the sum of each action's $\frac{c}{n}$ or $\frac{n}{c}$ for each participant. This sum was intended to represent the average degree of trust put into the chatbot, where a negative value would signify an opposite trust towards the newsfeed.

However, while considering the *ratio* between the two predictions, this index does not take the *difference* of the predictions into account. In other words, the analysis recognised the difference between predictions claiming large changes in value compared to those suggesting small changes in value. To address this, each $\frac{c}{n}$ or $\frac{n}{c}$ was weighted with the absolute value of the difference between two predictions:

$$t_d = t_r \cdot \left| p_{chatbot} - p_{newsfeed} \right|$$

where t_d is trust weighted with the difference in predicted values, t_r is either $\frac{c}{n}$ or $\frac{n}{c}$ and p is the predicted relative change of the portfolio.

There is one last aspect to consider. It is clear that a participant investing a greater amount in a portfolio displays more trust than a participant investing a smaller amount. To address this, the trust required of each action is weighted with amounts relative to the total wealth of the participant (available balance + invested balance) at the time of the action:

$$t = t_d \cdot \frac{a_{action}}{a_{available} + a_{invested}}$$

where t is trust required weighted with the action's amount (the final *trust index*), t_d is trust weighed with the ratio and difference of predictions and a are amounts.
Again, the sum of the value of t for each trading action can be obtained for each participant to quantify overall trust put into the chatbot (or the newsfeed) during the whole duration of the study, this time taking ratios of predicted values,

difference in predictions and the size of the action into account.
An alpha level of .05 was used for all statistical tests, which were performed using the statistical software *Jasp* (v0.10.2)[63].

4.4 Results

General Descriptive Results
In total, participants sent 1100 messages to the chatbot (M = 55.00, SD = 29.30) and submitted 1008 tag attempts (M = 50.40, SD = 57.81) over the length of the simulation, which was 20 minutes. The average final total balance was £1525.43 (SD = £479.08).

Overall, participants trusted the advice of the chatbot 115 times (M = 5.75, SD = 3.24) and the predictions of the newsfeed 72 times (M = 3.60, SD = 3.09) in contradictory situations, which participants faced a total of 187 times (M = 9.35, SD = 4.68). The average size of trading actions following the chatbot's advice was £116.91 (SD = £55.74), while the average size of actions following the prediction of the newsfeed was £116.97 (SD = £71.41). The average *trust index* towards the chatbot from participants was +11.24 (SD = 31.44).

On average, 3.2% (SD = 5.8%) of the messages were sent through the notification, while participants in the RS condition used reply suggestions 81.6% (SD = 10.0%) of the time.

NRS-RS Conditions (see Table 1, Table 2)
The advice of the chatbot was followed more often from participants in the RS condition (M = 78.3%, SD = 17.5%) than from participants in the NRS condition (M = 49.0%, SD = 19.8%) in situations of contradictory predictions. The difference is significant; $t(18) = 3.50$, $p = .003$.

The presence of reply suggestions had no significant effect on the size of the trading actions when following the advice of the chatbot; $t(18) = 1.90$, $p = .074$. However, there seems to be a trend with participants in the RS condition performing larger actions (M = £139.11, SD = £38.08) than participants in the NRS condition (M = £94.70, SD = £63.35).

When accounting for the size of the trading actions performed, the *ratio* and the *difference* between the predictions of the chatbot and the newsfeed, participants in the RS condition generated a higher *trust index* (M = +29.88, SD = 26.85) than participants in the NRS condition (M = -7.45, SD = 24.32). This difference is significant; $t(18) = 3.26$, $p = .004$.

	t	df	p
Chatbot follow ratio	3.499	18.000	0.003
Chatbot follow size	1.900	18.000	0.074
Trust index	3.258	18.000	0.004

Table 2. Independent samples t-tests

Regarding Image-Tagging, the difference between the number of tags generated by participants in the RS condition (M = 66.80, SD = 57.09) and the number of tags generated by participants in the NRS condition (M = 33.00, SD = 55.90) is not significant; $U = 30.50$, $p = .148$.

Questionnaire responses (see Table 3, Table 4)
Overall, in a Likert scale from 1 to 5, where 1 is the least amount and 5 is the most, participants claimed to trust the chatbot more (M = 3.45, SD = 0.95) than the newsfeed (M = 2.85, SD = 0.88) by the end of the study.

Results indicate that participants in the RP condition claimed to trust the chatbot more (M = 3.90, SD = 0.74) than participants in the NRP condition (M = 3.00, SD = 0.94) by the end of the study; $t(18) = 2.38$, $p = .029$.

On the other hand, there is no statistical difference between how much participants in the RP condition claimed to trust the newsfeed (M = 2.50, SD = 0.85) compared to participants in the NRP condition (M = 3.20, SD = 0.79) by the end of the study; $t(18) = 1.91$, $p = .072$.

Following the results of the Likert-scale questions are the participants' comments on the interaction with the chatbot and the platform.

	Chatbot follow ratio		Chatbot follow size		Trust index		Tag attempts	
	NRS	RS	NRS	RS	NRS	RS	NRS	RS
Mean	0.490	0.783	94.695	139.114	-7.436	29.883	33.000	67.800
Std. Deviation	0.198	0.175	63.351	38.076	24.323	26.848	55.895	57.088
Minimum	0.125	0.444	2.510	79.390	-61.905	2.270	0.000	0.000
Maximum	0.857	1.000	206.667	200.000	25.385	75.792	183.000	169.000

Table 1. Descriptive statistics of chatbot follow ratio, chatbot follow size, trust index and tag attempt

	Chatbot trust		Newsfeed trust	
	NRS	RS	NRS	RS
Mean	3.000	3.900	3.200	2.500
Std. Deviation	0.943	0.738	0.789	0.850
Minimum	2.000	3.000	2.000	1.000
Maximum	4.000	5.000	4.000	4.000

Table 3. Descriptive statistics of Likert-scale responses

	t	df	p
Chatbot trust	2.377	18.000	0.029
Newsfeed trust	1.909	18.000	0.072

Table 4. Independent samples t-test

35% of participants mentioned the word *accuracy* in their reply to the question *What would have made you trust the chatbot more?*, while 20% mentioned more personalised and human-like communication, and another 20% mentioned justification for the advice given.

Example of these answers are:

"Seeing that he was mostly accurate as the months passed by."

"If it was more personalised, i.e. if it used your name and varied its sentence structure."

"Some justification of the opinion that he holds."

To the prompt *Please leave your comments about your experience interacting with the Chatbot*, a particularly interesting remark was made from one participant:

"I didn't really trusted the Chabot, though I read all its suggestions. I could see that sometimes his predictions were wrong that's why I was hesitant to trust it. The interaction though was smooth"

Finally, comments on the overall experience with the platform were generally positive, but did not reveal additional indications of trust in the chatbot.

4.5 Discussion

Main Results

Supporting hypothesis *H1*, the main results suggest that the presence of reply suggestions in the interface of the chatbot encouraged participants to trust the chatbot more frequently. Because the situations considered featured contradictory statements between the agent's advice and the prediction of the newsfeed, this result implies that users tend to trust chatbots in a greater number of occasions if reply suggestions are visible.

An explanation of this result could take into account a possible interpretation of suggestions from participants. It can be argued that, because suggestions are embedded in the interface of the chatbot, they are assumed to be generated by the agent itself as a demonstration of its competence within the social trading environment. As research suggests that trust in competence affects benefit/risk-sharing efforts [19], it is likely that participants in the RA condition attributed a greater degree of responsibility for their trading actions to the agent.

Additional results reveal that, in the absence of reply suggestions, participants followed the chatbot's advice roughly as frequently as they followed the predictions of the newsfeed. Because the agent and the newsfeed phrased their predictions in an almost identical manner and provided the same amount of information, it could be argued that the interaction with a chatbot generates no inherent difference in trust compared to the acquisition of information from a newsfeed. However, this implication is not in line with previous findings relating to the effectiveness of risk communication, which suggest that people put more trust in *revealed* information (expressed in advice) than *stated* information (expressed in a source) [48]. In fact, while predictions were phrased in a similar manner, it was made clear to participants that the chatbot would offer *advice*, while the newsfeed would *state* information derived from external sources. Additionally, it is reasonable to assume that, because of the conversational nature of the interaction with a chatbot that displayed a basic degree of human likeness, participants were aware that the information offered by the agent consisted of *advice*. Nevertheless, it is possible that participants did not perceive this distinction to an adequate extent, in which case a more human-like chatbot offering *revealed* information in a different way could have produced different results.

H2 was rejected by results implying no effect of reply suggestions on the size of the trading actions when following the advice of the chatbot. Interestingly, despite trusting the agent more frequently, participants in the RA condition did not invest or withdraw greater amounts when doing so. It is important to remember that, in this condition, reply suggestions were shown at every stage of the conversation, including the moment participants were to determine the size of a trading action, where suggested

amounts were based on proportions of the available or invested balance, depending on the action. Given the fact that suggestions were used, on average, 81.6% of the time (*SD* = 10.0%), it could be argued that, if selected, suggested amounts were similar to the amounts specified by the participants in the NRS condition.

The average frequency of reply suggestions (81.6%) is interesting, as it can be compared to previous studies performed on Google's email application *Inbox* designed for mobile devices [28]. Here, *Smart Replies* were responsible for assisting with 10% of email replies. The difference in frequency of use is large. However, it is obvious that in one case, participants interacted with a chatbot, while in the other, users interacted with other users. Therefore, emails are likely to be composed more carefully, especially in a professional context where the degree of accuracy of suggestions is currently insufficient to replace manual typing.

Results indicate a significant difference in the *trust index* towards the chatbot displayed by participants in the RS and NRS conditions, where reply suggestions encouraged a greater index, in line with *H3*. As the index takes into account the context of each trading action, the study considers these results as a better indication of trust than results supporting *H1*. From this, it is fair to assume that not only did participants in the RA condition trust the chatbot more frequently than in the NRS condition, but they did so performing trading actions that were riskier overall. A possible explanation of this behaviour is that participants were less hesitant to perform an action if a suggestion was given, whereas the absence of suggestions encouraged participants to give more importance to the predictions of the newsfeed and to their available balance. Alternatively, it could be reasoned that participants using reply suggestions were less likely to acquire this information at all. In these cases, the advice of the chatbot was followed blindly, with participants displaying an extreme degree of trust towards the agent.

Of particular interest is the average *trust index* of participants in the NRS condition: -7.45 (*SD* = 24.32). Overall, despite following the advice of the chatbot roughly as frequently as the predictions of the newsfeed, this group of participants performed trading actions that displayed more trust towards the newsfeed. However, because some design aspects of the chatbot such as human-likeness were not explored fully in the study, it is difficult to infer that users trust chatbots with no reply suggestions to a smaller degree than they trust an external source of information like a newsfeed.

The final hypothesis *H4* was rejected by results indicating no significant difference between the amount of effort spent in the secondary Image-Tagging task by participants in the RS condition and participants in the NRS condition. Additionally, the distribution of tag attempts highly differed from normality and 5 participants attempted 0 tags. Therefore, a reasonable interpretation of these results needs to consider the possibility that the secondary task was unclear to some participants in the context of the study, a supposition supported by some of the questionnaire results, where 3 participants expressed confusion regarding the purpose of the task.

Finally, results regarding the questionnaire at the end of the online study reveal that participants in the RS condition claimed to trust the chatbot more than participants in the NRS condition. Here, it is crucial to recognise the difference between actual and self-reported degrees of trust, whose validity should be questioned in the face of self-report bias [15]. In fact, a general tendency for participants to present a favourable image of themselves on questionnaires [49] could imply that participants in this study preferred to report a more skeptical view of the chatbot (to avoid "getting fooled" by the study). However, regardless of social desirability bias, the statistical difference between the responses suggest that the self-reported degree of trust remained greater for participants with access to reply suggestions.

Applications

Possible applications of the findings of this study are numerous and varied, most of which relate to automation in customer service. Automation and online self-service solutions do not currently meet users' needs fully [16]. Meanwhile, trust has been shown to be a key factor in user uptake of interactive systems [7,20]. The present study suggests that the integration of reply suggestions in chatbots will increase user trust in the system and, consequently, increase their performance in the field of customer service. In this specific context, the study contributes to research revealing human-likeness, self-presentation and level of professional appearance as factors for trust in customer service chatbots [17].

An application of a chatbot of similar nature to the one developed in the scope of a social trading platform should also find the results of this study of particular interest. The study did not evaluate the effects of reply suggestions on trading performance. However, previous research suggests that, if a conversational agent gains the user's trust, then user satisfaction is enhanced [31]. Therefore, in the occasion of a possible integration of a conversational agent, an existing platform should consider integrating reply suggestions in the interface of the chatbot. However, platforms should observe that, if users are given the freedom to perform trading actions through the agent, the presence of reply suggestions has no effect of the size of

these actions. This is relevant, given that social trading platforms tend to generate revenue based on the spread they apply on each trading operation that is copied from each user [62].

Limitations and Improvements

There are several limitations that can be identified regarding the design of the online study and the analysis of the results. These concern aspects of validity, in particular internal and ecological validity.

Referring to the ability to draw confident causal conclusions from the research [33], internal validity ensures that results are robust and replicable [45]. Literature states that most of the threats to internal validity in experimental behavioural economics result from a common failure to assign subjects randomly to different treatments, while making comparison between treatment groups run at different times and with different population of subjects [33]. This does not apply to the present study, since the recruitment of participants ensured a random selection of a varied sample within the Prolific pool and a random assignment to conditions handled by the web application server. However, a different threat concerns the design of the simulated social trading platform. As the simulation was run over a discrete period of "months", participants were able to access information regarding the gain of each portfolio throughout the duration of the study. Therefore, the accuracy of the chatbot's advice could be clearly inferred by the actual change in value of portfolios at the end of each month. Research reveals that people place greater weight on information received from sources who have been more accurate in the past [14,21]. It follows that the accuracy of the chatbot's advice could have had potentially significant effects on the participants' trust in the agent. This variable was not controlled in order to design towards a random and unpredictable trading environment where the likelihood of participants recognising the mechanisms of actual portfolio gain generation was low. An improvement in this aspect would consist of keeping portfolio gain indicators hidden until the end of the study so that the accuracy of the chatbot's advice is unknown to participants. However, such an approach would have to consider different limitations regarding the resulting threats to the ecological validity of the study.

In experimental economics, ecological validity can be determined by the extent that the context in which subjects cast decisions is similar to the context of interest [42]. The design of the platform attempted to achieve this by replicating key features of existing social trading platforms, such as portfolio gain and risk indicators. However, since balances were necessarily virtual, it could be argued that the decisions made by the participants do not accurately reflect a realistic behaviour. In fact, while a bonus reward was given based on the final total balance, a minimum reward was set to guarantee that participants were fairly compensated for their time. An improved study would integrate the chatbot in an existing social trading platform to ensure minimal disturbance to the contextual ecology and, thus, produce more meaningful results [42].

5. CONCLUSION

The study explored the effects of reply suggestions on user trust in a conversational agent integrated in a simulated social trading platform. By developing a web application and conducting an online study on a sample (n = 20) of participants, it was possible to draw conclusions on the degree of trust displayed towards the agent by analysing the extent to which trading advice was followed. Results indicate that users with access to suggestions trust chatbots more frequently and perform riskier actions than participants with no access to suggestions.

Contribution was made to research on factors of trust in chatbots employed in the field of customer service, namely human-likeness, explanations and professional appearance. It is in this field that findings appear most promising, as user trust is proven to bring great benefits to both service performance and user satisfaction.

Future work could investigate the design aspects of reply suggestions, such as appearance order and content neutrality to determine whether these elicit more trust towards the agent. Additional studies could favour the use of autonomous agents over a social trading platform to determine the degree of autonomy that is delegated to the system as an alternative measure of trust. It would then be interesting to compare the findings with the ones presented in this study.

ACKNOWLEDGEMENTS

I am particularly grateful for the assistance given by my supervisor Dr Enrico Costanza and his invaluable contribution to the development of the research question, the acquisition of relevant literature and the deployment of the web application to the UCL server. I would also like to thank my family for displaying great interest in the study and for their many words of encouragement.

REFERENCES

1. Alberto Alesina, Eliana La Ferrara. 2000. The determinants of trust. NBER Working Paper 7621. National Bureau of Economic Research, Cambridge, MA, USA.
2. Manuel Ammann, Schaub Nic. 2016. Social interaction and investing: Evidence from an online

social trading network. https://www.rsm.nl/fileadmin/home/Department_of_Finance_VG5_/PAM2016/Final_Papers/Nic_Schaub.pdf.

3. Izak Benbasat, Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems*, 6(3): 72-101.
4. Philip C. Cai, Robert C. Miller, Carrie Jun Cai, James R. Glass. 2014. Wait-learning: Leveraging wait time for second language education. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, ACM, New York, NY, USA, 2239-2244.
5. Heloisa Candello, Claudio Pinhanez, and Flavio Figueiredo. 2017. Typefaces and the perception of humanness in natural language chatbots. *CHI '17 Proceedings of the 2017 CHI Conference On Human Factors In Computing Systems.* May 6-11, Denver, CO. ACM New York, NY, USA.
6. Minjee Chung, Eunju Ko, Heerim Joung, Sang Jin Kim. 2019. Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research.* Forthcoming.
7. Cynthia L. Corritore, Beverley Kracher, Susan Wiedenbeck. 2003. On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies* 58,6: 737–758.
8. Dianne Cyr, Khaled Hassanein, Milena Head, Alex Ivanov. 2007. The role of social presence in establishing loyalty in e-Service environments. *Interacting With Computers* 19,1: 43-56.
9. M.R. Delgado, R.H. Frank, E.A. Phelps. 2005. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience* 8,11: 1611-1618.
10. Daantje Derks, Agneta H. Fischer, Arjan E.R. Bos. 2008. The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior* 24,3: 766-785.
11. Philipp Doering, Susanne Neumann, Stephan R. Paul. 2015. A primer on social trading networks–institutional aspects and empirical evidence. https://pdfs.semanticscholar.org/1863/56c2b55289b61b0c85308ecfed9d4d3a538c.pdf
12. Stewart I. Donaldson, Elisa J. Grant-Vallone. 2002. Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology* 17,2: 245-262.
13. Derek Doran, Sarah Schulz, Tarek R. Besold. 2017. What does explainable AI really mean? *arXiv.org* > cs > *arXiv*: 1710.00794.
14. Ilan Fischer, Nigel Harvey. 1999. Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting* 15: 227–246.
15. Robert J. Fisher, James E. Katz. 2000. Social desirability bias and the validity of self-reported values. *Psychology & Marketing* 17,2: 105-120.
16. Asbjørn Følstad, Knut Kvale, Ida Maria Haugstveit. 2014. Customer support as a source of usability insight: Why users call support after visiting self-service websites. https://www.researchgate.net/publication/277309750_customer_support_as_a_source_of_usability_insight_why_users_call_support_after_visiting_self-service_websites
17. Asbjørn Følstad, Cecilie Bertnussen Nordheim, Cato Alexander Bjørkli. 2018. What makes users trust a chatbot for customer service? An exploratory interview study. *Internet Science*, Svetlana S. Bodrunova (ed.). Springer Nature Switzerland AG, Cham, Switzerland, 194-208.
18. Ulrich Gnewuch, Stefan Morana, Marc T.P. Adam, Alexander Maedche. 2018. Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. In *Proceedings of the European Conference on Information Systems (ECIS)*, Portsmouth, UK.
19. Byoung-Chun Ha, Yang-Kyu Park, Sungbin Cho. 2011. Suppliers' affective trust and trust in competency in buyers. *International Journal of Operations & Production Management* 31, 1: 56–77.
20. Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessy Y. Chen, Ewart J. de Visser, Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53,5: 517–527.
21. Nigel Harvey, Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes* 70: 117–130.
22. Matthew Henderson, et al. 2017. Efficient natural language response suggestion for smart reply. https://static.googleusercontent.com/media/research.google.com/it//pubs/archive/46057.pdf
23. Jennifer Hill, W. Randolph, Ingrid Farreras. 2015. Real conversations with artificial intelligence: A

comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49: 245-250.

24. Brett W. Israelsen and Nisar R. Ahmed. 2017. "Dave...I can assure you ...that it's going to be all right ..." -- A Definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *arXiv.org* > cs > *arXiv*:1711.03846
25. Jakob Nielsen, 2005. Ten usability heuristics. https://tfa.stanford.edu/download/TenUsabilityHeuristics.pdf
26. Devon S. Johnson, Kent Grayson. 2005. Cognitive and affective trust in service relationships. *Journal of Business Research* 58,4: 500-507.
27. Cynthia Johnson-George, Walter C. Swap. 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.
28. Anjuli Kannan, Peter Young, Vivek Ramavajjala, Karol Kurach, Sujith Ravi, Tobias Kaufmann, et al. 2016. Smart Reply. *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*
29. Sherrie Y. X. Komiak, Izak Benbasat. 2006. The Effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly* 30,4: 941.
30. Roderick M. Kramer. 2001. 1. Organizational paranoia: Origins and dynamics. *Research In Organizational Behavior* 23: 1-42.
31. SeoYoung Lee, Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103: 95-105.
32. Bingjie Liu, S. Shyam Sundar. 2018. Should machines express sympathy and empathy? Experiments with a health advice chatbot | *Cyberpsychology, Behavior, and Social Networking* 21,10: 625-636.
33. George Loewenstein. 1999. Experimental economics from the vantage-point of behavioural economics'. *The Economic Journal*, 109: F25–F34.
34. Baozhou Lu, Weiguo Fan, Mi Zhou. 2016. Social presence, trust, and social commerce purchase intention: An empirical research. *Computers in Human Behavior 56*: 225-237.
35. Fergus Lyon, Guido Möllering, Mark N.K. Saunders. 2011. *Handbook of Research Methods on Trust*. Edward Elgar Publishing Ltd., Northampton, MA, USA.
36. Inaki Maurtua, 2009. Wearable technology in automotive industry: From training to real production. In *Human-Computer Interaction*. Inaki Maurtua (Ed.). Available from http://www.intechopen.com/books/human-computer-interaction/wearable-technology-in-automotive-industryfrom-training-to-real-production.
37. Microsoft News Center India. 2018. BSE Launches an AI based Chatbot "Ask Motabhai" that Answers Stock Related Queries. 2018. Retrieved 7 August 2019, from https://news.microsoft.com/en-in/bse-launches-chatbot-ask-motabhai-for-faster-more-convenient-access-to-stock-market-information/
38. Christine Moorman, Gerald Zaltman, Rohit Deshpande. 1992. Relationship between providers and users of marketing research: the dynamics of trust within and between organizations. *Journal of Marketing Research* 29: 314-328.
39. Katie Delahaye Paine. 2003. *Guidelines for Measuring Trust in Organizations*. The Institute for Public Relations.
40. Wei Pan, Yaniv Altshuler, Alex Paul Pentland. 2012. Decoding Social Influence and the Wisdom of the Crowd in Financial Trading Network. *2012 International Conference On Privacy, Security, Risk And Trust And 2012 International Confernece On Social Computing*.
41. John K. Rempel, John G. Holmes, Mark P. Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49,1: 95-112.
42. Brian E. Roe, David R. Just. 2009. Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics* 91,5: 1266–1271.
43. Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint No.1708.08296*.
44. F. David Schoorman, Roger C. Mayer, James H. Davis. 2017. An Integrative Model of Organizational Trust: Past, Present, and Future. *Academy of Management Review* 32,2: 344-354.
45. Arthur Schram. 2005. Artificiality: The tension between internal and external validity in economic

experiments. *Journal of Economic Methodology* 12,2: 225-237.

46. Leila Takayama, Victoria Groom, Clifford Nass. 2009. I'm sorry, Dave: I'm afraid I won't do that: social aspects of human-agent conflict. *CHI '09 Proceedings of the 27th International Conference on Human Factors in Computing Systems,* Boston, MA, April 04-09. ACM, New York, NY, USA, 2099-2108.
47. Gábor Tatai, Annamária Csordás, Árpád Kiss, Attila Szaló, László Laufer. 2003. Happy chatbot, happy user. In: *Intelligent Virtual Agents*. T. Rist, R.S. Aylett, D. Ballin, J. Rickel (eds.). Springer, Berlin, Heidelberg.
48. Matt Twyman, Nigel Harvey, Clare Harries. 2008. Trust in motives, trust in competence: Separate factors determining the effectiveness of risk communication. *Judgment and Decision Making* 3,1: 111-120.
49. Thea F. Van de Mortel. 2008. Faking it: Social desirability response bias in self-report research. *Australian Journal of Adavanced Nursing* 25,4: 40-48.
50. Danding Wang, Qian Yang, Ashraf Abdul, A., Brian Y. Lim, 2019. Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference On Human Factors In Computing Systems* – CHI 2019, May 4-9, Glasgow, Scotland, UK. ACM, New York, NY, USA.
51. Vet Wolhlgemuth, Elisabeth S.C. Berger, Matthias Wenzel. 2016. More than just financial performance: Trusting investors in social traning. *Journal of Business Research* 69,11: 4970-4974.
52. Bo Xiao, Izak Benbasat. 2007. E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly* 31,1: 137-209.
53. Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, Rama Akkiraju. 2017. A new chatbot for customer service on social media. *CHI '17 Proceedings of the 2017 CHI Conference On Human Factors In Computing Systems*. Denver, CO. May 06 – 11. ACM, New York, NY, USA.
54. https://www.blog.google/products/gmail/save-time-with-smart-reply-in-gmail/
55. https://developers.facebook.com/docs/messenger-platform/send-messages/quick-replies/
56. https://www.djangoproject.com/
57. https://www.etoro.com/about/
58. https://www.etoro.com/discover/people
59. https://www.etoro.com/blog/trading-essentials/risk-score/
60. https://getbootstrap.com/
61. https://www.gov.uk/national-minimum-wage-rates
62. https://investingoal.com/how-does-etoro-make-money/
63. https://jasp-stats.org/
64. https://www.javascript.com/
65. https://jquery.com/
66. https://www.mysql.com/
67. https://www.prolific.co/
68. https://www.python.org/
69. https://rasa.com/
70. https://restfulapi.net/
71. https://www.sqlite.org/index.html
72. https://www.zulutrade.co.uk/