Investigating the Effect of Credibility Cues on Trust and Information Seeking in Chatbot Advisors

Nicole Soh
HCI-E MSc Final Project Report 2023
UCL Interaction Centre, University College London
Supervisor: Dr. Enrico Costanza

ABSTRACT

The incremental growth of chatbot advisors within investment contexts indicate the importance of understanding the influence of various design features in these scenarios. Literature reveals credibility to be an important facet influencing user trust in chatbots, thereby indicating importance in examining how it can be embedded into chatbot design features. Information processing theories also suggest individuals may subsequently base credibility judgments on superficial chatbot features. However, users may be motivated to investigate beyond these initial judgments if they do not sufficiently trust it. Pertinent literature findings indicate trust, information seeking, and credibility to be associated variables. Yet, these variables have not been tested together, nor have they been tested under more consequential settings, as prior studies have largely conducted them under non-experimental settings or scenarios involving low risk. The current study thereby examines credibility cues within chatbot recommendations in a simulated investment scenario, and its effect on information seeking and trust. A between-subjects experimental study with 22 participants investigated participant interaction with credibility-centric cues. Trust was quantitatively measured by how much participants trusted the chatbot over an alternative information source, under a simulated situation of risk where participants would be rewarded for their performance on the tasks. T-tests and correlational analyses of trust variables revealed no significant effect of credibility cues on interaction or trust towards the chatbot. Findings nonetheless contribute towards design guidelines of credibility cues on cognitively demanding interfaces such as those seen under investment trading scenarios, and the importance of considering user expectations in credibility-centric designs.

Author Keywords

chatbot; credibility cues; information seeking; interface design; user trust; social trading; online study

ACM Classification Keywords

H.5.m. Information interfaces and presentation: Miscellaneous

MSc Contribution Type

Empirical

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, 2023.

TABLE OF CONTENTS

1. INTRODUCTION

With the growth of artificial intelligence (AI), chatbots have in parallel become increasingly commonplace and relevant in everyday usage [25].

Chatbot implementations can range from informal communication—such as the layperson interactions with Siri or Alexa—to more formal communications within the workplace. Given the flexible implementations of such systems, organisations have more recently started to look towards implementing chatbots due to the increased productivity such bots bring towards the organisations [39]. Here, chatbots provide can provide workers with the convenience and support towards addressing queries they may have– which help organisations minimize the resources required to tend to menial user requests. The numerous ways a chatbot can be trained to interact with its users also mean that it can help with more complex decision-making processes. One industry in which the use of chatbots have become increasingly popular is the financial sector. Financial institutions have begun to implement robo-advisors to provide financial advice to customers who can then choose to act upon these advices. Such robo-advisory systems have been measured to be advantageous in affordability towards the organisations that provide such services, in that they are widely scalable and cheaper to maintain, than human advisors [51]. A study by Mesbah et al. [31] examined the perceived advantages of AI-based advisories– such as those of robo-advisors– compared to human experts. Participants reported a perceived advantage to using robo-advisors compared to human advisors, which included perceptions of robo-advisors being convenient to use as a primary advantage, while reliability and objectivity of the robo-advisors were seen as a secondary advantage. *Statistica* report that as of April 2023, there are an estimate of 183.70 million users who have reported using such systems to invest their money [47]. This is estimated to grow up to 234.30 million users in 2027. The rapid growth within the robo-advisory market alongside the rate at which financial organisations uptake such systems, indicate that understanding the factors influencing system adoption is a lucrative point of focus, as these factors can directly impact whether the projected usage of such chatbots is met and even exceeded. There are challenges in light of this projected growth, as numerous aspects of the chatbot interface can influence whether users decide to adopt the system. It is important that organisations understand these factors influencing system uptake, as it will ensure that developments towards implementing a chatbot is not futile.

While there are numerous reasons to why users choose to adopt chatbot usage, one prominent factor consistently associated with the uptake of such robo-advisors is whether users hold sufficient trust in the system [31] This is supported by a systematic literature review of 51 studies [57] finding that trust was frequently positively associated with the direct intention to use chatbots—the greater the trust users have towards the chatbot, the more likely they are to use it. Such trust is thereby important to establish, as it directly supports the hypothesized growth of the robo-advisory market. There are several factors that contribute towards user trust in chatbots. This include cues which the chatbot provides to the user, such as visual appearance, conversational approach, or the identity type of the chatbot [14]. Per Ejdys [10], trust in technology (e.g. chatbots) may also associate with factors such as security, credibility, reliability, loyalty, and its performance accuracy. In this, design is integral in ensuring that indicators of trust are integrated effectively into the chatbot platform, as the aforementioned factors need to be shown to users through manner of speech, information presentation, or explanation. These indicate ways which a chatbot can be designed to foster user trust towards it, not restricted to outward, superficial designs. It can also involve the manner which the chatbot is designed to present critical and more meaningful information towards users—including how users are able to access such information.

Across research, a particularly important relationship between trust and credibility has been identified to have a significant impact on whether technological projects are able to be successfully implemented [56]. Tseng and Fogg [52] proposed that achieving credibility is imperative under certain human-computer interactions, some of them being when computers: provide users with data (1), report measurements (2), and when reporting on work performed (3). Importantly, these interactions frequently occur in robo-advisory/financial contexts. This suggests that achieving credibility alongside trust is particularly important, given that there will be risk associated with trusting financial chatbots to implement their advice on the users' behalf. The present study thereby aims to examine whether the design of chatbot interfaces can incorporate credibility elements which influence user trust and behaviour under conditions of risk.

Altogether, these emphasize the importance of understanding how credibility can be incorporated into the design of chatbot systems to improve user trust. The current study thereby aims to contribute to pre-existing literature by examining the influence of credibility-centric interface designs on user trust and information seeking behaviours. To understand how to best incorporate credibility into design and measure its effects, a literature review was conducted on research within the human computer interaction (HCI) and cognitive psychology fields, where findings were used

to inform the current study's measure of trust, and support the development of a credibility-centric cue. These designs were tested through a between-subjects study design, to examine whether the presence of credibility cues induce feelings and behaviours of trust under an experimental scenario. Findings are evaluated against prior literature, providing implications on effective design of credibility cues within investment trading interfaces.

2. LITERATURE REVIEW

2.1 Trust

Trust within the HCI literature widely points towards an individual's confidence in a computer agent's ability to be depended on, reliable, and functioning as expected by the user [41].

A greater level of trust is associated with numerous benefits, for example-- increased user trust is associated with increased provision of user data towards algorithms, where the increased data is used by the algorithm to produce higher quality outcomes; and allowing systems to function as intended by the organisation [45]. There is a low value yield for users in situations where they do not trust the recommendations provided by the chatbot [44]. In this, establishing user trust is particularly important for chatbots in the financial sectors, as these systems provide advice and recommendations for users to act upon.

Initial trust in a technology may also indicate how an individual subsequently interacts with it. Kim and Benbasat [19] conducted a literature review on information systems, identifying that a prominent way of increasing trusting beliefs is to allow users to interact and examine sources. In studies of online recommendation agents, initial trust from the initial use of the system will subsequently influence likelihood of adopting use of the system [58]. These findings were similarly found by Benbasat and Wang [1], where the initial trust of participants were associated with the perceived usefulness and intentions of using a recommender agent.

Nonetheless, the literature debates on how trust is measured—complicating the process of a cohesive definition of trust across research. Glaeser et al. [13] find that attitude measurements of trust only weakly predict trusting behaviours, and posit that prior trusting behaviours will predict future trust better. Such findings posit that trust is best measured through behaviour, aligning with the manner which Milana et al. [33] conceptualized and measured trust. The study by Milana et al. involved a situation of risk, where participants undergoing an investment scenario made decisions on recommendations by two sources that contradicted each other—one source being a chatbot, the other being a newsfeed providing an alternative source of information. Outcomes of these recommendations would not be known to users until after a certain amount of time passed, meaning that users would have to choose one source over the other. Choosing to follow the recommendations of one source would thereby indicate that the participant trusted that source over the other in a condition of uncertainty. The scenario by which Milana et al. invokes trust thereby comprises of a deterministic decision rather than a subjective measure of trust i.e. attitude measures [33]. Likewise, this measure of trust takes trust into consideration under conditions of uncertainty. This measure of trust can thereby be argued to be more applicable to situations involving risks— much of which can be observed in the financial sector and when making investments [4]. While users in Benbasat and Wang's [1] study exhibited intentions to use the agent more if they trusted it, it is important to note that the measure of trust intention in their study involved minimal risk, as users were not required to act on recommendations provided. Using the scenario similar to Milana et al.'s study provides the current study with an opportunity to examine how initial trust translates into intentions or behaviours, in a situation where risk is involved. In support of this, McKnight et al. [30] also indicate that trusting intentions can be defined by whether the individual is willingness to depend, provide information, make purchases, and following the advice of the source. The current study therefore deploys the same scenario by Milana et al. [33] to measure trust, as the manner which trust is measured aligns broadly with the manner that trust intentions is defined by McKnight et al. [30].

2.2 Credibility

Across literature, researchers have arrived at the consensus that credibility refers to the perceived credibility of the technology, rather than the objective credibility. Tseng and Fogg [52] thereby defines credibility as whether the technology is believed to be able to produce quality outcomes. This aligns with the source credibility theory (SCT) [27], which posit that if individuals perceive a product to be credible, they are subsequently persuaded more effectively by it. This in turn has outcomes on how the source is seen as trustworthy and reliable— which a large body of research uncovered that credibility generally comprises of two key components: trustworthiness and expertise [46]. Thus, when a technology is considered credible, it can be implied that it is found to be both trustworthy and of expertise. Credibility is

important to achieve, as it is one of the three main concepts associated with trust [7].

Studies have found various means by which technology can achieve credibility. Fogg et al. [12] indicate one of the most applicable channels of doing so is through surface credibility. Surface credibility refers to the quick initial judgments made on "surface" features of the product, such as logos or labels, which lead users to form impressions of the product's credibility [50, 34]. Koh and Sundar [21] examined the effect of adding labels when describing websites, finding that labelling websites/technological systems with a specifying label– i.e. "Wine website", "wine computer", "wine agent"– contributed towards user perception that these systems were subject specialists. This fostered an increased sense of trust, perception of expertise, and purchase intention from these users, highlighting the significant influence of a simple labelling effect. Websites including third party endorsements are associated with a lower level of perceived risk of purchasing from the website [18] Websites with a privacy seal associate with an increased provision of data by participants [37]; and websites which incorporate more of such design features were associated with greater perceived credibility [38].

Logo designs that specifically incorporate elements of credibility also have been found to associate with larger interaction rates [16], emphasizing the impact of credibility-centric designs. This was further supported by research of Lowry et al. [27] which implemented credibility traits into logo design—i.e. traits reflecting trustworthiness, expertise, and an additional dimension of dynamism, which is the way in which credibility is confidently communicated by the source. Lowry et al. [27] reasoned that dynamism was included due to the large body of research evidence indicating that it is a key factor in effectively portraying credibility through a source. Fogg et al. [12] revealed dynamism to play an important role in influencing the perception of surface credibility, with their participants indicating that the presentation of information strongly influenced their perceptions of credibility. Lowry et al.'s [27] findings provide support that logos specifically designed with the aforementioned credibility traits are positively associated with user trust, and can influence subsequent behaviour. This provides indication that cues specifically integrated with these credibility traits may also positively associate with user trust and subsequent behaviour.

Successful implementations of surface credibility-centric chatbot design have also been reported. Liew et al. [50] found that chatbots integrated with expertise cues—labels indicating them to be product specialists—were perceived more as experts and similarly, more credible, than chatbots that did not have these cues present. Likewise, participants who interacted with the expertise chatbots exhibited greater intentions to purchase items from the chatbot.

Altogether, research on surface credibility indicate that simple cues can have a significant influence on the perceived credibility of the system. These point towards a minimal cue effect [55], that people tend to rely on superficial/little information to form an impression within an uncertain situation. Nonetheless, the studies on credibility largely measured only intentions, but not the actual behaviour. Intentions may not necessarily translate into behaviour, thereby highlighting the need to examine whether users who interact with a chatbot containing credibility cues would exhibit trusting intentions as those defined by McKnight et al. [30]. Likewise, studies on credible chatbot designs focus on increasing credibility through the implementation of expertise cues into the appearance and communication style of the chatbot [50]. In this, there seems to be limited studies assessing simple design cues with specifically embedded credibility traits—as had been done by Lowry et al. [27]– into chatbot information design. Dynamism has not been widely incorporated into credibility-centric chatbot designs. Likewise, in a survey of over 2,500 respondents, information design/structure was said to be important in the formation of perceived credibility [12]. These imply the importance of considering credibility cues within the information structure of chatbot recommendations. Supported by Benner et al. [2] finding small design elements to be successful in the context of product recommendations, the presentation of cues which establish chatbot recommendations/advice as "credible" or "believable" may provide interesting insights which have not yet been revealed in prior studies, within an investment scenario.

2.3 Information Processing

Iterated by SCT, the initial impressions users have of the product can be critical in influencing decision to further interact with the product [e.g. 11, 26]. It is thereby important to understand how information processing of interfaces occur, as this provides insights to how users identify aspects of credibility across the vast array of information available on a technological interface.

While there may be numerous features of an interface that influence user trust, people have been proposed to have an information processing capacity [22] This

means that there is a limit to how much users can process at any given time, where not all aspects of the system will be processed by the user. Specific aspects of the system, rather than the whole system, will be taken into account when influencing user trust. Sundar [49] argues that many web users rely on visual cues involving texts, images, or logos to make quick judgments about information quality—in order to reduce the time and cognitive cost of processing the vast array of information occurring at any given time on an interface [49]. This indicates it is important to understand the impact of design elements within interfaces, as these elements may be meaningfully processed by the user and can subsequently shape their impressions on trust and perceived credibility of the interface.

The use of heuristics is proposed to explain how individuals choose specific information, where heuristics have been identified to be frequently engaged in persuasive communication design [2]. Heuristics are mental shortcuts used to identify important information in the environment. This way, the time and cognitive load it takes to process information in the environment is significantly reduced [32]. However, individuals may also choose to search for specific information beyond the immediately noticeable information available on the interface. This can occur when there is a motivation to do so, and/or when they are cognitively able to. Iterated, it is argued individuals who seek for information will more carefully evaluate the arguments presented on websites, as they are motivated by a goal [38]. This aligns with the Elaboration Likelihood Model.

The Elaboration Likelihood Model (ELM) proposes that individuals' attitudes are changed through two main routes: the central and peripheral route [36]. The peripheral route refers to superficial cues, which relate to how surface-level credibility can be achieved by quick initial glances at websites or interfaces [27]. Users might be inclined to follow this route when they are not cognitively able to process richer, in-depth cues such as meaningful messages. Arguments/meaningful content pertain to the more central route of processing information, where the meaning of messages are more carefully evaluated in order for the user to be persuaded [36]. Numerous studies have indicated that under situations with limited time to engage in decision making [40], heuristic processing of peripheral cues play an important role in influencing user attitudes. Heuristic processing is suggested to occur frequently in investment scenarios due to uncertainties, volatility, and imperfect nature of the information provided within these situations [4]—highlighting the importance of superficial cues in shaping user decision making. This is supported by studies examining the user experience of retail investors when using online investment platforms, finding that due to the time pressure within such scenarios, many investors adopt decision rules—i.e. heuristic processing, allowing them to make decisions in time limited situations [24].

The effectiveness of peripheral cues have been reported to be more significant than that of central cues (i.e. message content) in such time-limited scenarios, one of such peripheral cue being source credibility [3, 28]. In this, the heuristic being invoked is that "credible sources are trustworthy", leading users to be persuaded by the heuristic which they enable. Chaiken and Maheswaran [3] examined the influence of credibility cues on information processing, revealing that when participants considered a task to be of low importance, the heuristic processing of such cues were the sole determinants of attitude despite more elaborate, persuasive messages being presented alongside the credibility cues. Nonetheless, when individuals are motivated to search beyond the initially presented cues of a website/interface, they are likelier to carefully evaluate beyond these cues [38]. Research by Metzger et al. [32] also reveal that the processing and evaluation of central cues—such as those of message content, is likelier across users who are highly motivated to determine if the source is trustworthy. This motivation has been found to associate with initial trust.

Several studies on consumer mobile health behaviour found that trust was a factor that influenced information seeking [5, 9], suggesting that lower trust encourages information seeking, while higher trust decreases the need to information seek. This is further supported by research from Shin [45]. Trust was implicated as a key role in influencing whether users perceive an algorithm to be credible. The survey study examined the interaction between trust, information seeking, and perceived credibility of a chatbot journalist that provided recommendations. Their finding revealed that trust was significantly associate with information seeking, where information seeking occurred to determine the chatbot's credibility. In turn, information seeking/interacting more with the chatbot was found to be associated with increased trust—users indicated that they were more willing to share their data with the chatbot. Findings by Le [23] support this relationship, finding that interactivity with a chatbot in their study was associated with increased adoption intentions of the system. Nicholas et al. [35] also evaluated information seeking behaviour in online environments, with the study uncovering that users

engaged in rapid cross-checking behaviour between web contents. One reason they posit this behaviour occurs is due to the vast information widely available, which lead users to quickly check between sources to determine which one they can trust.

Such findings imply trust and information seeking are a bidirectional relationship, which presents an opportunity for the current study to examine this relationship within an environment involving financial risk.

3. CURRENT STUDY

Numerous associations between trust, credibility, and information processing are present across pre-existing research. Likewise, there are areas which have not yet been explored in greater detail. Much of prior research on chatbot design and surface credibility focused on incorporating credibility cues into chatbot appearances and communication types [50] but have not focused widely on incorporating design cues into the information structure of chatbots, such as into the advice they produce. Many studies also do not measure trust outcomes/trust intentions through behavioural measures—iterated in the literature review on trust measures. While the measurement of trust intentions in Milana et al. [33] align with definitions of trust intentions [30], the study focused on understanding the effect of reply suggestions, and not perceptions of credibility. Iterated, trust and credibility are widely related associations [7]. This presents an opportunity for the present study to examine how surface credibility influences trust intentions in a situation characterized by risk, similar to that of Milana et al. [33]. This would contribute insights on whether incorporating credibility cues into chatbot designs are sufficient to influence not just intentions, but tangible behaviours of trust intentions. Understanding how information seeking occurs under conditions with surface credibility can also provide insights on how interfaces can be shaped to accommodate such behaviour.

3.1 Research Aims and Hypotheses

The current study thereby presents several hypotheses in light of research reviewed.

Per measurement of trusts by Milana et al. [33] and per findings within the credibility literature [27]:

H1: Participants exposed to credibility cues will follow chatbot advice more frequently than participants not exposed to credibility cues.

H2: Participants exposed to credibility cues will make on average, a larger action size when following the chatbot's advice, compared to participants not exposed to credibility cues.

H3: Participants exposed to credibility cues will exhibit a greater chatbot trust compared to participants not exposed to credibility cues.

Per the literature on ELM, trust, and information seeking [23, 32, 44]:

H4a: Participants who have a lower baseline trust in technology will evaluate credibility cues of the chatbot more frequently.

H4b: Participants who more frequently evaluate credibility cues will have higher perceived trust in the chatbot.

3.2 Design Rationale

The study uses the same online investment simulation per Milana et al. [33], largely maintaining the same interface structure. This interface comprises of two sources presented side-by-side, a recommender "assistant" chatbot and a newsfeed that provides predictions; and a section where users can keep track of the amounts of each virtual portfolio on the investment simulation. Users make investment actions by communicating with the chatbot through messaging, where the actions they can take are to follow, unfollow, increase amount invested in, or decrease amount invested in a particular portfolio. Users may also seek advice from the chatbot, where the chatbot will provide recommendations on who to follow, unfollow, and who to invest more or less in.

While Milana et al. [33] deployed four variations of the chatbot, the present study decided that only one chatbot condition would be used. We chose to not delay the response of the chatbot in this study, as varying responses may contribute to users' perception of the chatbot being humanlike, which may in turn influence their perception of the chatbot [20]. The interaction of these alongside the variable we plan to manipulate in our current study may lead to an interaction effect that we were not intending to measure, thereby going beyond the scope of the present study. Likewise, reply suggestion buttons were not included for this current study, as it may influence participants' perception of trust within this study and potentially have an interaction effect with our credibility cues. To understand the influence of credibility cues on user trust, we thereby used the default state of the chatbot, which did not contain variable speech or reply suggestion buttons.

Placement Of Credibility Cue

Milana et al. [33] justified the design for their chatbot recommendations to include minimal data, as disclosing exhaustive data was suggested to overwhelm participants who do not have sufficient trading knowledge to make use of such data. However, the present study decided to embed the credibility cue into each recommendation provided across the chatbot and the newsfeed. This was done as the cue would not comprise of specialist knowledge which would disadvantage low-knowledge users.

Credibility Cue Design

Design iterations were created through Figma [59], before one design was implemented into the simulated investment scenario through code, for pilot testers to interact with.

The credibility cue was designed to engage a credibility heuristic [49], which would act as the cue users saw on each recommendation. This was done by specifically incorporating design guidelines found to be associated with surface credibility [27, 18]. Initial ways the credibility cue could be displayed was explored, shown in Figure 1.

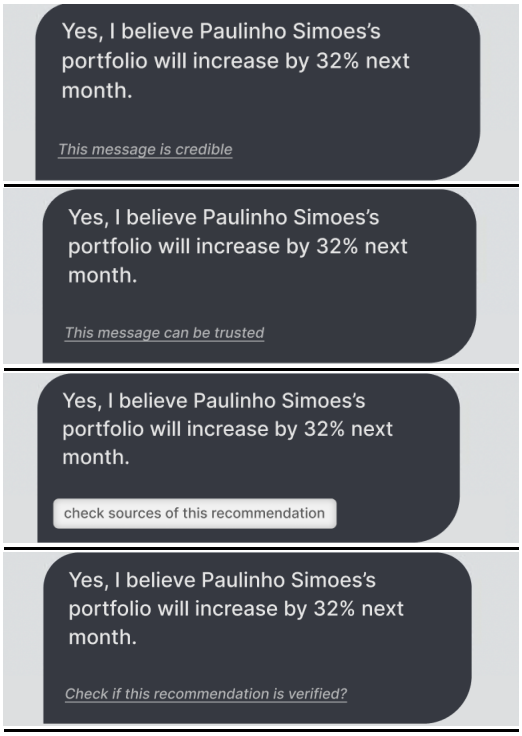

Figure 1. Initial iterations of credibility cues on the chatbot's interface were explored through Figma

For the pilot test design, each recommendation was attached with a cue stating: "recommendation is verified". This is supported by common perceptions of credibility on social media being found to associate with the term "verified" [54] where this was subsequently included in our design of the credibility cue, due to the common associations of credibility with being verified. This phrasing also incorporated dynamism [27] into the cue, as "being verified" implies confidence in the source being "correct/believed".

To measure whether participants evaluate the credibility cue— thereby indicating information seeking behaviour, the cue can be interacted with (clicked on) to generate a more elaborate message. The message generated would pertain to endorsements— as prior studies found the addition of such information to significantly associate with perceived credibility [18]. These would be implemented in the form of the recommendation aligning with independent sources. 6 variations of the message content were created and randomized across recommendations, with each variation stating how credible the specific recommendation was. This was added as a form of message variability to serve as a control of ecological validity. This would range from no source supporting the recommendation, to multiple sources supporting it. The layout for these variations would be the same, apart from the text content elaborating how credible these recommendations were. Independent sources would endorse the message, as endorsement by third-parties were found to be associated with increased perceived credibility [18].

While the message content for each recommendation was randomized across portfolios, the message content across the newsfeed and chatbot was fixed. This meant that if a recommendation for a specific portfolio in the newsfeed had four sources aligning with it, the chatbot's recommendation for the same portfolio would also be expected to have four sources that aligned with it. This was done to ensure that participants did not choose the chatbot over the newsfeed because it had more sources aligning with it compared to the newsfeed—or vice versa. We could thereby infer that participants who choose one source over the other, do so due to their perceived trust or credibility for the source, rather than the objective differences in credibility levels between sources.

While it might be assumed that a recommendation which more sources aligned with would be a more accurate recommendation, it was decided that the number of sources aligning with a recommendation would not specifically associate with an intended outcome. This was done due to the random, volatile nature of investment scenarios [4]. Likewise, the present study aimed to examine perceived credibility, rather than objective credibility, on associated behaviours. Thus, changes in portfolio amounts occur

randomly, and are not contingent on the number of sources aligning with a recommendation.

Credibility cues were also integrated across both chatbot and newsfeed to ensure that participants were not biased to prefer one source over the other due to data availability—i.e. participants preferring the chatbot because it showed information that sources aligned with it, while the newsfeed did not show this information. By integrating cues across both sources, participants trusting one source over the other would indicate that an effect on perceived credibility, rather than objective credibility between source, was observed.

3.3 Pilot Testing

The initial design was pilot-tested with 5 individuals recruited through the researcher's personal network, age (M=24.4, Min= 22, Max= 26).

These designs were high-fidelity prototypes of the actual simulation that the real participants would interact with, and were shown to pilot testers through the researcher's laptop where the investment scenario was set up locally. Pilot testers interacted with the system for one simulation month, to test viability of the design presentations for the real study. They were then asked to provide feedback on their perceptions of the credibility cue, including the message content when it was clicked on.

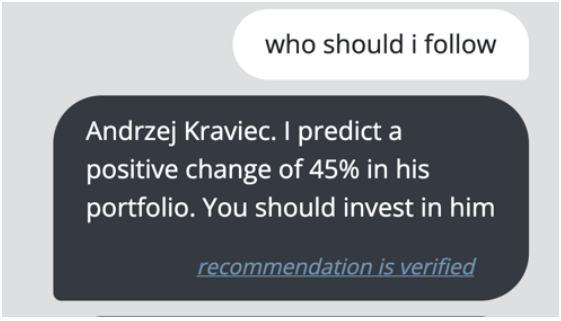

Figure 2. The surface credibility cue shown to pilot testers.

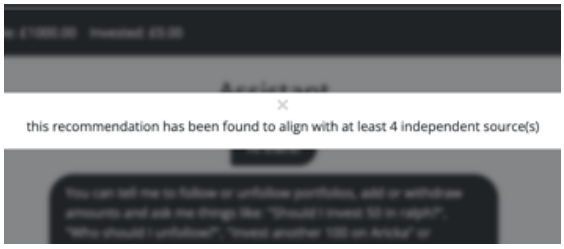

Figure 3. The content shown in the pilot-test when the credibility cue is interacted with.

The design of the credibility cue and its content presented to pilot testers is shown in figure 2 and 3.

This design was featured across the newsfeed and chatbot. Feedback was then gathered about the design presentation, and content wording when the credibility cue was interacted with. 3 of the pilot testers feedbacked that the phrase "recommendation is verified" was too long, and made the interface look visually overwhelming especially when they had to process information quickly with a limited timer. This led to a design revision, which involved shortening the phrase to "verified", alongside an addition of an inquiry logo which was included to indicate to users that the cue was interactive and could be clicked on for more information.

For the content within the credibility cue, pilot testers indicated that the phrasing "independent sources" did not exhibit sufficient relevance to the financial recommendations given by the newsfeed or chatbot, suggesting that there should be more specificity to the phrasing—such as the source being relevant to the investment scenario. This included suggestions to include specific real world recognisably-expertise sources—e.g. "the Financial Times". While this was not incorporated into the final revision due to potential confounds of participants having prior exposures and differing opinions of real-world sources, the phrasing was revised to indicate these were independent financial sources. This was supported by prior research finding no difference in perceived authority between a fictitious and real authority cue [49]. Thus, to avoid the possibility of participant bias, a non-descript authority source was used—"independent financial source". The final design (shown in Figure 4) incorporated feedback from participants, and was implemented through both front-end and back-end code modifications, elaborated in the development process.

Development Process.

To make the proposed design changes into the pre-existing interface used by Milana et al. [33], the source code was kindly provided by Dr. Enrico Costanza and Federico Milana. The major changes for the current study involved front-end additions of credibility cues attached to all newsfeed and chatbot recommendations. *HTML* and *CSS* code were modified to incorporate the design and layout format of the credibility cues. *Python* was used to modify pre-existing bot messages to append a credibility cue to all advice given by the chatbot, either when advice was automatically generated, or in response to an advice request made by a participant. This was similarly done to all recommendations appearing on the newsfeed. Python code pertaining to speech variability and reply suggestion buttons were ignored for the conditions of the current study. The JavaScript code was also

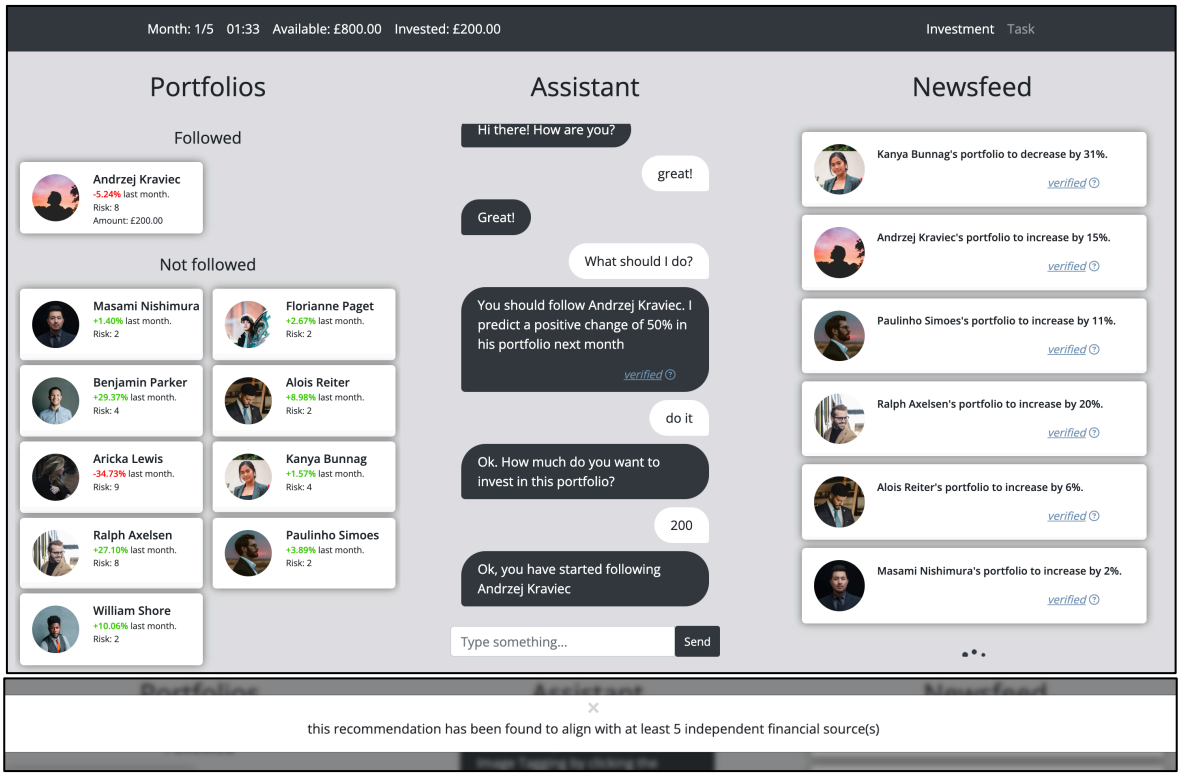

Figure 4. The finalized credibility cues and message content shown to the study's participants

modified to allow the randomization of message content across portfolios, whilst remaining constant between the chatbot and newsfeed. Back-end Python code was also adjusted to ensure that interaction clicks were stored on the database. These were subsequently tested to ensure that participant interactions were correctly recorded across the chatbot and the newsfeed. Once tested, the revised application was uploaded onto the UCL server by Dr. Enrico Costanza.

Altogether, these modifications allowed the present study to carry out its aims of examining the influence of credibility cues on user behaviours within an investment scenario.

3.4 METHOD

Materials/Stimuli

The current study used the investment scenario by Milana et al. [33], with modifications set out in the design rationale. All participants will be given £1000 of virtual money to use throughout the duration of the investment scenario.

To measure initial trust of participants, 3 items pertaining to user trust in technology from the Social Service Robot Interaction Trust (SSRI) scale was used [6]. These items focused on users' initial trust stance on technology– found by McKnight et al. [29] to be significant in influencing user behaviour towards technology. These items where thereby used to measure user trust in AI before starting the investment scenario, where the composite score of the three items will be taken as the user's initial trust score.

An image-tagging task is used as the secondary distractor task, which participants are allowed to engage with if they wished to during the investment scenario. This was done to increase the ecological validity of the study, as trading within the real world often involves individuals who do not focus on only one task, but may include multiple sub-tasks. Successfully engaging the image-tagging task each time would increase the virtual amount that participants had. To successfully complete each tagging task, participants were required to provide 3 word tags associated with the image they were shown. When all word tags were identified for each image, users would be rewarded with an £20 of in-study currency. This amount was determined by Milana et al. [33], to be an appropriate pay-out per image tagged. If participants were unable to find all 3 word tags attached with one image, they were able to skip to the next image task.

Attention check measures were added across the questionnaires. This was done to minimize nonsense responding, and to ensure that the data used for analyses did not contain careless responding.

Participants

Participants were recruited from Prolific, an online platform that provides participants for research studies [62]. A total of 22 participants completed the study, ranging 19 – 36 years old (M=25.7, SD= 4.90). Across the participants, 12 identified as a Man (Including trans-man/male), 9 identified as a Woman (Including trans-female/woman), and 1 identified as non-binary. 15 Participants had no experience with investment, while 7 participants had experience. Majority of the participants were of white ethnicity (N=16), the remaining participants were of black ethnicity (N=6). Most participants were from Poland (N=5), and South Africa (N=5). Participants were screened to ensure they did not complete the Milana et al. [33] investment chatbot study, and had a minimum age of 18 due to Prolific's minimum age criteria. There were no other exclusion criteria.

On average, participants completed the study in 30.53 minutes (SD= 4.89, Min= 24.78, Max= 47.13). Participants were paid for their participation in the study, where the study operated on an incentive basis. For participating, participants would be paid £5. This roughly translated to the rate of £10 per hour, aligning with the standard of the UK national minimum wage as of April 2023 [60]. Participants were given the opportunity to earn incentive bonus capped at £3, where £1 was awarded for each in-study currency of £300 earned during the investment simulation.

Study Design

A between-subjects design (control vs. credibility) was used to address the research hypotheses. Participants were randomly allocated to either the control condition or the credibility condition. Allocation was done by codes implemented into the server, which ensured that both conditions would comprise of equal participants.

Participants in the control condition would not be presented with credibility cues. Participants in the credibility condition would be shown credibility cues embedded onto each recommendation, across the chatbot and newsfeed. In the credibility condition, participants would also be able to interact with these cues.

The independent variables manipulated is the presence of the credibility cue, and the dependent variables measured is the chatbot follow ratio, average trading action size following the chatbot's advice, and trust towards the chatbot, including perceived trust.

Procedure

The online study was conducted on Prolific over the span of one workday, which participants accessed through their personal computers. The duration of the study lasted for 30 minutes. This consists of the time participants would take to complete the investment scenario and questionnaire. Participants underwent the same procedure across the control and credibility conditions. The only differences were the manipulations of the independent variables on the interfaces across the different conditions; and differences on instructions of how to interact with the interfaces across conditions.

Participants were first briefed about the study. This consisted of the study's purpose, duration to complete all expected tasks, the reward, and how participant data will be collected and stored. Participants were also informed that they could withdraw from the study at any time, and where to direct questions they might have about the study. Once briefed, participants were directed to give their consent at an informed consent form page. Once consent was provided, participants were first tasked to complete the initial questionnaire on AI trust, as adapted from Chi et al. [6]. Upon submitting the first questionnaire, participants proceed to the main section of the study. The main section of the study comprised of a 20 minute interaction with the simulation scenario. Participants could choose to switch between the investment task and the image-tagging task, with the goal of earning in-study currency. Tasks can be switched by toggling between the "investment" and "task" tab on the notification bar at the top of the screen. The 20 minute simulation was split into 5 months. Each month would last for a total of 4 minutes, before advice given by the chatbot and newsfeed would refresh—where both sources would produce a newly randomized set of recommendations. A randomized set of message contents attached to credibility cues were also refreshed each month, for participants in the credibility condition.

In the first month, participants are greeted by the chatbot assistant, which provides instructions on how to interact with it and what it can do. As each month ends, participants are informed that a new month has started and that portfolio values have changed. The change in values affect all portfolios, including the ones which participants choose to invest in and the ones that participants are not following. Once the 20 minutes are up, participants will be redirected to a results summary page, which provides information on the total amount of virtual money earned throughout 20 minute simulation. They would then proceed to the post-study questionnaire. Across the conditions, there are 5 Likert-scale items and 4 open questions, none were optional. The first two items pertained to the perceived trust of participants towards the chatbot assistant and newsfeed, and the fourth and fifth item pertained to the perceived credibility of participants to

the chatbot assistant and newsfeed. The open ended questions inquired about what could increase participants' trust towards the assistant, their experiences interacting with it, and the overall experience of the study. Once participants provide their input for all items, they were then redirected to Prolific to confirm that they have completed the study.

Data Processing and Cleaning

All data processed passed both attention check measures (100%). All participants in their respective conditions (Control, N= 11; Experimental, N= 11) were included for data analysis of the following variables—initial trust in technology; perceived trust in the chatbot and newsfeed; number of cue interactions.

Similar to Milana et al. [33], to analyze follow ratio, average action size, and trust index of the chatbot across the conditions, data was filtered to only record actions made in the instances where participants chose to perform an action during contradictory recommendations. Contradictory recommendations occur when the chatbot and newsfeed provide recommendations that occur in the opposing direction. Processing only these points of action would allow us to identify the source which participants preferred in situations where an absolute choice had to be made, if the participant chose to perform an action—either trusting the chatbot's recommendation, or the newsfeed recommendation. This means that there might be contradictory instances which participant did not make an action. This would then not be recorded in the database. Two participants (control participant 56 and experimental participant 73) did not make any actions during contradictory moments. They were thereby indicated to make 0 actions during contradictory instances. For the t-tests ran, these data were excluded from analysis comparing across the two conditions. This was decided as both datapoints made equivalently 0 actions, and were each from the experimental and control condition respectively. This thereby meant that removal of the datapoints would not bias results towards one condition over the other. 10 participants each from the control and experimental condition were subsequently ran for analysis of mean difference in: *chatbot follow ratio, average action size when following the chatbot, and trust in the chatbot* (H1, H2, H3).

User interactions with the credibility cues were measured across all recommendations, instead of during only contradictory recommendations, as the aim of H4a and H4b was to measure user interaction with credibility cues and whether participants' perceived trust would be associated with information seeking behaviour. Cue interactions were thereby collected across all recommendations provided, for each source.

Statistical Analysis Methods

To address H3, the current study measured trust for each action made, with the *trust index* formula devised by Milana et al.:

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

This *trust index* formula [33] ensured that the trust participants exhibited by choosing the chatbot over the newsfeed's recommendations— t, accounted for the dynamic changes in variables during the investment scenario. More specifically, the investment scenario involved dynamic prediction changes across the newsfeed and chatbot. For example, the newsfeed may propose Portfolio X to decrease 77%, but the chatbot contradictorily proposed Portfolio X to increase 3%. Given the disproportionate predicted difference, it is likely that participants would choose to follow the newsfeed's recommendation because the newsfeed's predicted change was much larger. This was accounted for through t_d, which is the trust participants displayed towards the chatbot/newsfeed weighted with the difference in predicted changes between the chatbot and newsfeed. Participants who invest a larger proportion of their existing balance at the time of action can also be seen to display more trust towards the chatbot—but this value may differ depending how much each participant had at the point of each action. This was accounted for by considering the ratio between the amount participants invested (a_{action}) and the total balance ($a_{available} + a_{already\ invested}$) that they had at the time of action. This devised the trust index formula by weighting the magnitude difference between the chatbot and newsfeed recommendations t_d, and the relative amounts participants had at each action made.

To address H4, the perceived trust variable is measured by the composite score of participants' perceived credibility and trust in the source, based on prior research indicating them to be widely related [7].

Statistical outliers identified were retained in the dataset, as data was collected from a representative sample through Prolific. Outlier data were thereby assumed to be natural variation of the population. Additionally, not all responses of the outlier participants were deemed as statistical extreme,

suggesting that outlier responses for each portion of the study is due to natural variation. It was thereby determined that all outliers identified would be retained. However, this will be stated in the result analysis wherever identified.

The data obtained was checked to ensure that assumptions for the different tests were met. To examine the differences between the credibility and control conditions, an independent samples t-test would be used– as the means of two unrelated groups would be compared. As data was gathered across the conditions separately, it can be assumed that independence of the observations was met. Likewise, the dependent variables measured were of ordinal nature – the number of times advice was followed, and of continuous nature – the ratio of times the chatbot followed the chatbot over the newsfeed, and the chatbot trust index. Extreme outliers were tested by running a boxplot diagram for the conditions. One outlier was detected for the action size variable (Participant 54)**.** Three outliers were detected (Participant 40, 43, and 61) for the trust index variable. Homogeneity of data was met, with all dependent variables scoring non-significant *p* values. Normality of data used was checked with a Shapiro-Wilk test—where tests of Normality returned significant p values for the *chatbot action size* variable*,* and the *trust index* variable. Thus, a Mann-Whitney U test was used as an alternative to the independent samples t-test for these variables, as it does not require normality of data. For chatbot follow ratio and all questionnaire items, student's independent samples t-tests were ran, as these variables met the assumptions of normality across both conditions.

To test for an association between cue interaction and perceived trust variables in the credibility condition, spearman's correlation analysis was chosen as a non-parametric alternative to Pearson's analysis. This was because the normality of data was violated, $p < 0.05$, and outliers were present in the data for cue interactions. This would have significantly influenced the Pearson's correlation output—as it was determined that outliers would be ran across all analyses.

Data was cleaned through Microsoft Excel and analyzed with SPSS (v. 29.0) [61]. All statistical tests were analyzed with an alpha level of 0.05.

4. RESULTS

4.1 Quantitative Analysis

Descriptive Statistics

On average, the 22 participants scored 3.61 (*SD*=0.781) on their initial trust towards technology across conditions. Participants sent the chatbot a total of 847 messages (*M*= 38.50, *SD*= 13.54). In total, 1154 images were tagged (*M*= 52.45) across the 20 minute duration of the investment scenario. On average, participants had a final total balance of £1312.06 (*SD*= 477.91).

For the 20 participants that performed one or more action when the chatbot and newsfeed provided contradictory responses, a total of 169 contradictory situations (*M* = 8.45, *SD*= 3.56) were faced. Within these situations, participants followed the chatbot's advice a total of 94 times (*M*= 4.70, *SD*= 1.90), where the average size of trading action for the chatbot was 151.38 (*SD*= 104.97). Participants followed the newsfeed's predictions a total of 75 times (*M*= 3.75, *SD*= 2.90), where the average size of trading action was 145.38 (*SD*= 95.74). Participants had an average chatbot trust index of -8.79 (*SD*= 30.24).

Across the 11 participants in the credibility condition, participants clicked on the chatbot's credibility cues a total of 47 times (*M*=4.27, *SD*=5.23). The newsfeed credibility cues were interacted with a total of 73 times (M= 6.64, SD=7.35). Table 1 breaks down the descriptive statistics of cue interactions between chatbot and newsfeed.

	Credibility Cue Interaction	
	Chatbot	Newsfeed
Mean	4.27	6.64
Std. Deviation	5.26	7.35
Minimum	0	0
Maximum	16	18

Table 1. Descriptive Statistic of credibility cue interactions across the sources, for the credibility condition.

Experimental-Control Conditions

An independent samples T-test (and non-parametric alternative, Mann-Whitney U test) was run to compare the differences in average mean across the credibility and control conditions for H1, H2, and H3. Across the groups (N=20), no significant difference was found in the average chatbot follow ratio between the control

	Chatbot follow ratio		Chatbot action size		Trust Index	
	Control	Credibility	Control	Credibility	Control	Credibility
Mean	0.6644	0.5041	128.46	174.31	-1.45	-16.14
Std. Deviation	0.1903	0.1695	84.37	122.37	8.93	41.60
Minimum	0.3888	0.2500	13.64	67.25	-24.03	-131.30
Maximum	1.000	0.8000	268.19	427.79	11.86	14.96

Table 2. Descriptive Statistics of Variables Compared Between the Control and Credibility Condition

condition (*M*= 66.44%, *SD*=19.03%) and the credibility condition (*M*= 50.41%, *SD*= 16.95%), $t(18)= 1.99$, $p= 0.06$. There was no significant difference in the average action size of participants when following the chatbot's advice between the control (*M*= 128.46, *SD*= 84.37) and the credibility group (*M*= 174.31, *SD*= 122.37), $U = 58.00$, $p = .545$. The average trust index of the control condition (*M*= -1.44, *SD*= 8.93) was higher than the credibility condition (*M*=-16.14, *SD*= 41.60). However, this difference was non-significant, $U = 35.00$, $p = .257$.

Correlation Analyses

To test H4a and 4b on whether there was a relationship between trust and information seeking, i.e. the number of times credibility cues were interacted with, a spearman's correlational analysis was ran on data from the credibility condition.

Spearman's correlation analysis revealed a non-significant weak negative relationship between the initial average trust in technology (*M*=3.61, *SD*= 0.70) and number of times credibility links were clicked on chatbot recommendations (*M*=4.27, *SD*=5.26) for the 11 participants within the credibility condition, $r(9) = -.31$, $p = .347$. While non-significant, direction of results indicate that a low initial trust in technology score is weakly associated with a higher number of cue interactions.

To examine whether information seeking during the simulation was associated with perceived trust towards the chatbot at the end of the study, another spearman's analysis was ran. The analysis revealed a non-significant weak positive association between number of cue interactions and subsequent trust in the chatbot, $r(9) = .13$, $p = .701$. While non-significant, direction of results indicate that an increase in number of link clicks is weakly associated with greater perceived trust in the chatbot.

Questionnaire responses

By the end of the study, the 22 participants on average trusted the chatbot (*M*= 3.43, *SD* = 0.92) more than the newsfeed (*M*= 3.04, *SD*= 0.94). Independent samples t-tests were ran to determine whether there was a mean difference between the credibility and control conditions. Participants in the control condition had a greater perception of the chatbot (including trust and credibility) (*M*=3.55, *SD*=0.93) than in the credibility condition at the end of the study (*M*=3.32, *SD*= 0.93), but this difference was non-significant, $t(20)= .57$, $p = .574$. Participants in the control condition had a greater perception of the newsfeed (*M*=3.32, *SD*=0.78) than in the credibility condition (*M*= 2.77, *SD*= 1.03) by the end of the study, but this difference was non-significant, $t(20) = 1.40$, $p = .178$.

To examine whether the differences in the perceived trust towards the chatbot and newsfeed was

	Initial Trust AI		Chatbot Perceived Trust		Newsfeed Perceived Trust	
	Control	Credibility	Control	Credibility	Control	Credibility
Mean	3.61	3.61	3.55	3.32	3.32	2.77
Std. Deviation	0.892	0.696	0.934	0.929	0.783	1.034
Minimum	2.33	2.33	1.50	2.00	2.50	1.00
Maximum	5.00	5.00	5.00	5.00	5.00	4.50

Table 3. Descriptive Statistics of Trust Perceptions across conditions

significantly different between the two groups, an independent t-test was ran on the difference score (perceived trust in chatbot – perceived trust in newsfeed) between the two groups. Participants in the credibility condition (M= 0.55, SD= 1.27) did not significantly differ in difference score from the control condition (M= 0.23, SD= 1.54), $t(20)$= -.53, p= 0.603.

4.2 Qualitative Analysis

To analyze the qualitative data, a simple thematic analysis was conducted. This was done by summarizing each of the 22 participants' responses to the open-ended questions into codes, where common patterns were identified within these codes. When patterns were identified, a theme was formed. A description for each theme was developed, where responses which identified most suitably with these descriptions would be fit into those themes. These themes will be elaborated in the sections below, alongside example responses.

For the question "what would make you trust the assistant more", four themes were identified: Accuracy, Understandability, Interactivity, and Explainable Information.

Accuracy

The theme which the most responses fit into was accuracy (N=9, 40.90%). This theme was defined as the chatbot's correctness in its predictions over the newsfeed, or vice versa, which influenced participants' trust towards it. Example responses included: "Because he [chatbot] predicted the increases well".

Understandability

Responses within the understandability theme were defined to focus on understanding typing errors or queries which the participants had. 6 Participants (27.3%) indicated that an increased understanding from the chatbot would lead them to trust it more. Example responses included: "If only he understood me more" (Participant 73). The chatbot conversation of some of the participants whose responses fell under the understandability theme, was further analyzed. This was taken from the UCL server where data of the bot messages were stored. Some participants asked the assistant questions related to the tagging tasks, an example chat extract from Participant 73 to the chatbot: "How to enter tags?", where this request was made to the chatbot several times. This might have contributed to the participant's response to this question, as the chatbot was not programmed to focus or help participants with the tagging task, only the investment task. Likewise, responses from participant 59 ("If the AI could understand me better") included complex enquiries to the chatbot which it was not programmed to respond to, such as: "Can I move my investments from another account", and a subsequent clarification message: "Can I remove my investment from [name] and invest in [name]". Instead, the chatbot was programmed to only understand simpler messages that were broken down into separate inquiries. Such conversations may have contributed to responses falling within the understandability theme.

Explainable Information

6 responses (27.3%) were found to fit under explainable information theme. This theme was defined by ways in which the chatbot could have explained its predictions to users through elaboration, data, or design presentation. Participant 48 stated: "If the assistant gave a probabilistic outcome rather than an all out prediction", and participant 60 stating: "adding chances to the result they were suggesting.", provided insights that chatbot recommendations could have been designed to include descriptives or numerical values to support each recommendation given. Elaboration of explainable information involved the chatbot elaborating more on their predictions—"more comments" (Participant 61); "about how it predicted the changes […]" (Participant 51); or presenting their responses more simply (Participant 47), which suggests that the complexity of the chatbot's message delivery may have contributed to participants' perceived trust in it. Participant 42 stated "evidence that showed that what he suggested could really bring me profit"—suggesting that the presentation of factual data might have contributed to perceived trust in the chatbot.

Interactivity

Responses which fit into this theme consisted of descriptions about ways in which the chatbot could have been interacted with, which included interaction complexity and flexibility. Only one participant (4.55%) indicated that having the chatbot be more interactive would lead them to trust it more (Participant 56).

For the question asking participants to describe their experience interacting with the assistant, responses fell into three existing themes—accuracy (N=3, 13.64%), understandability (N=5, 22.73%), and interactivity (N= 8, 36.36%). A thematic analysis

revealed that data also fell into two new themes: verifiability and convenience. One participant stated a NULL response for this question, and was not included in abovementioned themes.

Verifiability

This theme was defined as the ability to verify the chatbot's recommendations which influenced participants' perceptions of trust and credibility in it. Two responses (9.09%) fit into the verifiability theme. Participant 57 indicated: "it became more credible to me when the investments would go up", suggesting that participants verified the chatbot's credibility through its observed successful predictions. Participant 58 indicated "I trusted the assistant more than any source although I still had to cross-check the predictions with other independent sources". This suggests that some participants still felt the need to information seek even though they trusted the chatbot, indicating that information seeking might play a role in the process of developing trust with the agent.

Convenience

This theme was defined as the benefit which the chatbot provided to the participant. Three responses (13.64%) fit into the convenience theme. This included the quickness/slowness of the chatbot in responding (Participants 56, 61); and the affordance which the chatbot provided for the participant to focus on other tasks, Participant 71 stating: "I could focus on doing the task while assistant was suggesting me what should I do to invest".

4.3 DISCUSSION

The current study examined user interactions and perceptions of credibility cues, and how the presence of such cues influenced trust in chatbots. However, none of the hypotheses proposed were supported. For H1, the average chatbot follow ratio was found to be higher for the control compared to the credibility condition, although this difference is non-significant. For H2, while the average action size is higher in the credibility condition, this difference is non-significant. For H3, findings contradicted the hypothesized direction, with participants in the control condition trusting the chatbot more than in the credibility condition. For H4a and 4b, trust and information seeking behaviours were non-significant, but weakly associated in the hypothesized direction. Qualitative analyses also revealed several factors which may have contributed to the present findings.

The rejection of H1 and H3 implied that the presence of credibility cues did not influence the participants to trust it more than when there were no credibility cues present. Interestingly, although non-significant, participants had a greater chatbot follow ratio in the control condition than when they were exposed to credibility cues— going against the hypothesised direction and prior findings [21, 50]. One possible explanation to such findings may be due to the accumulation of several factors. While the newsfeed cleared out each month, the chatbot kept all messages sent by the participant in the chat. This might have allowed participants to refer to previous messages and assess the accuracy of the chatbot, while not being able to do the same for the newsfeed. This might have led participants to form associations between the chatbot's predictions and its accuracy—where the violation of expectations between the chatbot's predictions and portfolio changes might have influenced trust participants had towards the chatbot. This effect may have been greater for participants exposed to the credibility cues, as the violation of expectation may have been accumulated through numerous cues associated with the chatbot—i.e. the inaccurate recommendation being listed next to the actual changes in portfolio in the subsequent months; the "verified" credibility cue associated with the inaccurate recommendation; and numerous independent sources which aligned with the inaccurate recommendation provided by the chatbot. The combination of these cues might have led to a greater expectation violation for participants within the credibility condition than in the control condition, which may explain the lower follow ratio and trust index for the chatbot in the credibility condition. Such findings are supported by Grimes et al. [15] indicating that users combine numerous cues of chatbot systems to form expectations of how the interaction would occur. Their study revealed that having high expectations for AI systems, including chatbots, to function in a particular manner and not having these expectations met, lead users to more negatively evaluate the agent, compared to when there is a low expectation of the system to function well and having it subsequently exceed expectations. Participants in their study were found to evaluate a system more critically when it violated positive expectations—where users criticized the system for performing even beyond its capabilities. This may have occurred for our study, particularly with the phrasing used for the credibility cue which was used to express dynamism [27]. It is thereby plausible there was a greater expectation for the chatbot to be more accurate when credibility cues were present, compared to when there was none—explaining the rejections of H1 and H3. A more neutral phrasing to prompt participants to interact with the cue can be considered, to reduce the likelihood of an extreme expectation violation.

Alternatively, user interviews can be conducted to probe user perceptions of chatbot expectations, and to better understand how user evaluate expectation violations of these chatbots, when it occurs.

H2 was rejected, as results were non-significant. The nonsignificant difference in trading action size for the chatbot in the credibility condition may be in part due to the low trust participants displayed towards the chatbot, per rejections of H1 and H3. Iterated, trust has been found to associate with user data provision [44]. In the present study's findings, it can be argued that as participants did not display more trust towards the chatbot, they thereby did not want to provide more "data"—the equivalent of investing beyond an average amount of money towards the chatbot's recommendations than they would have for recommendations of the newsfeed.

H4a and 4b were similarly rejected due to non-significance, although associations were in the hypothesized direction. An explanation to why the hypothesized relationships did not reach significance may be due to the design of the credibility cue, its' contents, and the process for participants to access the central cue. As the credibility cue was designed to display dynamism [27] in that its' credibility was believable, users may have evaluated this aspect of the cue—rather than engage in further information seeking by interacting with the cue. Rieh [42] posit that predictive and evaluative judgments continuously occur until the user stops searching. This was further supported by Unkel and Haas [53] not finding an effect between prior user characteristics and the preference to interact with information linked with credibility cues, where it was suggested that their participants' initial impressions of credibility may have inhibited their need to engage in more evaluative judgments of credibility, such as through interaction. Their findings similarly aligned with results attained for our study. In our study's case, searching to evaluate information beyond the initial "verified" credibility cue may have ceased once participants formed an initial judgment of credibility on the cue, thereby explaining the findings of H4a.

Likewise, the intended central cue was not immediately present. Participants had to navigate around each recommendation to access this. Iterated by Lang [22], humans have an information processing capacity. This was supported by Liao et al. [24] finding that participants making investments on an investment platform relied on heuristic processing to make their decisions. In a time-limited situation such as our study's investment scenario, participants with a low initial trust in technology may have been motivated to evaluate credibility arguments more carefully. However, the steps required to access the central cue may have demanded cognitive capacities which participants did not have, due to the need for participants to concurrently process multiple other aspects of the interface during the study—i.e. the newsfeed, the portfolio changes, the image tagging task. Thus, while users may be motivated to evaluate information—per the ELM [36]— barriers on the interface may have prevented these motivations from actualizing. The message content may also not have contained messages that were persuasive enough for participants to process it as a central cue. This was supported by the qualitative responses. Numerous participants within the credibility condition indicated seeking out other specific explainable information, such as "if the assistant gave probabilistic outcomes", "more comments", and "Evidence that showed that what he suggested could really bring me profit.". These suggest that the information which some participants were seeking might be incongruent with the information presented by the central cue, and altogether may contribute towards the non-significant results of H4b.

Implications

While hypotheses within the present study were rejected, the findings nonetheless provide implications for information presentation on chatbot interfaces. Per results, not all participants chose to information seek even when the option to was present. This provides an implication about user behaviour, in that users may not seek information even when information is available. Information that is intended to be conveyed should be readily accessible across sources—meaning that users should have access to important information about chatbot recommendation credibility as easily as possible. This can subsequently reduce the cognitive load it would have taken to seek out relevant information, minimizing information processing barriers for the user [22]. Iterated, expectation violation can lead to overcritical evaluations of the system [15]. It should thereby also be ensured that relevant information, when directly presented, do not contribute to unreasonable expectations beyond what AI systems can do—i.e. it should not overpromise on what it can do, but rather provide a metric that more reasonably indicates its ability.

Despite the nonsignificant results, findings from participant responses nonetheless highlight that participants engaged in information seeking across the newsfeed, chatbot, and independent sources—highlighted by responses such as " […] At the end I trusted the assistant more than any source, although I still had to cross-check the predictions with other

independent sources". This indicated that cross-checking with other resources may still be an important process of user interactions with chatbots on a complex interface—indicated by the participant's need to cross-check predictions, and as supported by Rieh [42] that users continuously engage in information seeking until they determine they have attained sufficient information. This provides implications that such an affordance to cross-check information provided by AI agents, should be provided when users interact with an interface. As highlighted by prior research [44, 23], this increased trust can in turn lead to increased adoption of the system.

Limitations and Future Improvements

Findings from the study should be carefully interpreted, as it cannot be completely assumed that participants cross-checked both newsfeed and chatbot sources before performing an action. This might be the case for some users, as mentioned across many users within the open-ended questions: "it predicted increases well. However, I wanted to make sure of my investment by using the newsfeed", and " if their suggestions aligned with the newsfeed". However, other responses such as "I trusted the assistant quite a lot at the beginning since I did not pay attention to the newsfeed at times" highlight that it is possible that not all participants cross-checked both sources before performing a user action on a contradictory recommendation. It may be worth performing an eye-tracking study on the present study's set up, as this may indicate how information processing occurs across the different elements of the investment interface. Likewise, it can provide support on whether users are actively cross-checking across the sources before making an informed decision, or whether the action they performed was triggered by the most recent information available to them—either that of a newsfeed post or chatbot's advice showing up.

The current study also implemented message content variation to increase ecological validity within the study. However, the effects of viewing the different variations were not disentangled. In this, participants may have viewed a message variation indicating that the recommendation aligned with numerous sources. This may have been the first or last piece of information processed/evaluated by the participant, which may have been impressionable in shaping participants' perceived credibility of the chatbot—aligning with the *primacy effect*, where information presented first is remembered best; or the *recency effect*, where information presented most recently/last is remembered best [8]. This is particularly relevant, as Steiner et al. [48] found that the performance evaluations were affected by how recently good/bad information was presented to the observer. This thereby indicates that such effects may have contributed to the results, highlighting the need for future studies to improve upon this aspect— i.e. through counterbalancing of the message variations.

Numerous participants also pointed out that the chatbot was unable to understand them sufficiently, as evaluated within the qualitative analysis. This ranged from minor typing errors, to enquiring about topics which the chatbot was not programmed to understand or respond to. Nonetheless, understandability plays a significant role on influencing trust in AI [43] and may thereby explain some of the findings associated with participants' trust towards the chatbot. This variable was not explicitly measured within the present study, making it difficult to disentangle from findings attained. Future studies can incorporate understandability as a variable, and examine through multiple linear regression, the extent to which it predicts trust compared to the other variables being assessed on the study.

Lastly, only a limited number of participants (N=22) could be recruited for the study due to budget constraints for this MSc project. These may have contributed to the false rejection of a true positive—i.e. a Type II error, as small sample sizes have been found to contribute to such errors [17] However, the hypotheses may have been plausibly rejected due to a genuine absence of an effect, reflecting similar non-significant findings by Unkel and Haas [53]. Nonetheless, there is value in replicating this study with a larger sample size, as this would reveal whether the hypotheses proposed were true null hypotheses, or whether they were consequent of a Type II error.

Future Work

The current study's rationale for including credibility cues across both sources was to ensure a non-biased perception between the chatbot and the newsfeed, as the cues contained data about varying sources aligning with the recommendation. Future work may want to explore the incorporation of a non-interactive credibility cue exclusive to the chatbot within the experimental condition—as done by Milana et al. on the incorporation of chatbot-specific buttons within their study [33]. This may allow future studies to examine whether the presence of credibility-centric

cues alone can contribute to greater perceived credibility in the chatbot.

Likewise, future study might want to explore the inclusion of other indicators within message variations. While the current study focused on one template of message content and credibility cue design, our qualitative analyses revealed that presenting credibility cues in alternative formats may lead to different findings. The explainable information theme revealed that further elaboration on the message content, or providing evidential and numeric outcomes would have led users to trust it more. These highlight a possibility for future studies to explore different ways to deliver messaging content to users, and their effect on perceived credibility. For example, it may be of interest to explore user behaviour towards central cues that are presented immediately on an interface, and to examine whether users choose to evaluate them meaningfully when barriers to accessing these cues are reduced. Alternatively, displaying the actual likelihood of the recommendation occurring as a statistic to the user, which substitutes the "verified" credibility cue within the current study may lead to different findings. It would be interesting to examine whether presenting objective magnitude-based cues would produce a more significant effect on perceived credibility.

5. CONCLUSION

Altogether, the current study examined the effects of surface credibility on information seeking and trust in a chatbot, under situations involving financial risk. This was tested with a between-subjects design involving 22 participants who used a simulated trading platform containing a chatbot embedded with a credibility-centric cue. T-tests and correlational analyses on participant interaction with credibility cues revealed that findings did not reach statistical significance, indicating that presence of the credibility cue did not significantly affect trust or information seeking as hypothesized. These findings nonetheless provided insights into design guidelines for presenting information within a demanding environment, such as that of a financial investment platform. Firstly, findings going against the hypothesized directions imply that the design of credibility cues should not over-inflate user expectations, as this may contribute to a harsher evaluation of the system when user expectations are not met. Secondly, important information should be easily accessible to users, particularly in an environment that is cognitively demanding. These findings also indicate areas where future research can be done, such as exploring alternative implementations of surface credibility on chatbot recommendations and examining its influence on user trust.

ACKNOWLEDGEMENTS

A big thank you to Dr. Enrico Costanza who has provided his invaluable time throughout the dissertation in giving feedback, technical support, and assistance for code implementations required for the project to run. I also show much appreciation to family and friends who have provided unconditional support throughout my dissertation journey.

REFERENCES

1. Izak Benbasat, and Weiquan Wang. 2005. "Trust in and adoption of online recommendation agents." *Journal of the association for information systems* 6.3: 4.
2. Dennis Benner, Sofia Schöbel, and Andreas Janson. 2021. "Exploring the state-of-the-art of persuasive design for smart personal assistants." *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*. Springer International Publishing, 2021.
3. Shelly Chaiken, and Durairaj Maheswaran. 1994. "Heuristic processing can bias systematic processing: effects of source credibility, argument ambiguity, and task importance on attitude judgment." *Journal of personality and social psychology* 66.3: 460.
4. Sayan Chaudhry, and Chinmay Kulkarni. 2021. "Design patterns of investing apps and their effects on investing behaviors." *Designing interactive systems conference 2021*.
5. Juan Chen, Xiaorong Hou, and Wenlong Zhao. 2016. "Research on the model of consumer health information seeking behavior via social media." *International Journal of Communications, Network and System Sciences* 9.8: 326-337.
6. Oscar Hengxuan Chi, et al. 2021. "Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery." *Computers in Human Behavior* 118: 106700..
7. Cynthia L Corritore., Beverly Kracher, and Susan Wiedenbeck. 2003. "On-line trust: concepts, evolving themes, a model." *International journal of human-computer studies* 58.6: 737-758.
8. William D. Crano. 1997. "Primacy versus recency in retention of information and opinion change." *The Journal of Social Psychology* 101.1: 87-96.
9. Zhaohua Deng , Shan Liu, and Oliver Hinz. 2015. "The health information seeking and usage behavior intention of Chinese consumers through mobile phones." *Information Technology & People* 28.2: 405-423.

10. Joanna Ejdys. 2018. Building technology trust in ICT application at a University. International Journal of Emerging Markets, vol. 13, no 5, pp. 980-997.
11. Andrea Everard, and Dennis F. Galletta. 2005. "How presentation flaws affect perceived site quality, trust, and intention to purchase from an online store." *Journal of management information systems* 22.3: 56-95.
12. Brian J Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. 2003. "How do users evaluate the credibility of Web sites? A study with over 2,500 participants." In *Proceedings of the 2003 conference on Designing for user experiences*, pp. 1-15.
13. Edward L. Glaeser, David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter. 2000. "Measuring trust." *The quarterly journal of economics* 115, no. 3: 811-846.
14. Eun Go, and S. Shyam Sundar. 2019. "Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions." *Computers in Human Behavior* 97 (2019): 304-316.
15. Mark G. Grimes, Ryan M. Schuetzler, and Justin Scott Giboney. 2021. "Mental models and expectation violations in conversational AI interactions." *Decision Support Systems* 144: 113515.
16. William Haig. 2007. "How and why credibility-based company logos are effective in marketing communication in persuading customers to take action: A multiple case study toward a better understanding of creativity in branding." PhD diss., Southern Cross University.
17. Luke J Harmon, and Jonathan B. Losos. 2005. "The effect of intraspecific sample size on type I and type II error rates in comparative studies." *Evolution* 59.12: 2705-2710.
18. Dan J Kim, Donald L. Ferrin, and H. Raghav Rao. 2008. "A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents." *Decision support systems* 44, no. 2: 544-564..
19. Dongmin Kim , and Izak Benbasat. 2003. "Trust-related arguments in internet stores: A framework for evaluation." *J. Electron. Commer. Res.* 4.2: 49-64.
20. Lorenz Cuno Klopfenstein , Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. "The rise of bots: A survey of conversational interfaces, patterns, and paradigms." In *Proceedings of the 2017 conference on designing interactive systems*, pp. 555-565.
21. Yoon Jeon Koh, and S. Shyam Sundar. 2010. "Effects of specialization in computers, web sites, and web agents on e-commerce trust." *International journal of human-computer studies* 68.12: 899-912.
22. Annie Lang. 2000. The limited capacity model of mediated message processing. *Journal of communication*, *50*(1), 46-70.
23. Xuan Cu Le. 2023. Inducing AI-powered chatbot use for customer purchase: the role of information value and innovative technology. Journal of Systems and Information Technology.
24. Li Liao, Zhengwei Wang, Jia Xiang, Hongjun Yan, and Jun Yang. 2021. User interface and firsthand experience in retail investing. *The Review of Financial Studies*, *34*(9), 4486-4523.
25. Abbas Saliimi Lokman, and Mohamed Ariff Ameedeen. 2019. Modern chatbot systems: A technical review. In *Proceedings of the Future Technologies Conference (FTC) 2018: Volume 2* (pp. 1012-1023). Springer International Publishing.
26. Paul Benjamin Lowry, Anthony Vance, Greg Moody, Bryan Beckman, and Aaron Read. 2008. Explaining and predicting the impact of branding alliances and web sitequality on initial consumer trust of e-commerce web sites. Journal of Management Information Systems, 24, 199–224.
27. Paul Benjamin Lowry, David W. Wilson, and William L. Haig. 2014. A picture is worth a thousand words: Source credibility theory applied to logo and website design for heightened credibility and consumer trust. International Journal of Human-Computer Interaction, 30(1), 63-93.
28. Durairaj Maheswaran, and Shelly Chaiken. 1991. Promoting systematic processing in low-motivation settings: Effect of incongruent information on processing and judgment. *Journal of personality and social psychology*, *61*(1), 13.
29. Harrison D. Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS), 2(2), 1-25. Chicago*
30. Harrison D. McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. Information Systems Research, 13, 334 – 359.
31. Neda Mesbah, Christoph Tauchert, Christian Michael Olt, and Peter Buxmann. 2019. Promoting trust in AI-based expert systems.
32. Miriam J. Metzger, Andrew J. Flanagin, and Ryan B. Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication*, *60*(3), 413-439.
33. Federico Milana, Enrico Costanza, and Joel E. Fischer. 2023, July. Chatbots as Advisers: the Effects of Response Variability and Reply Suggestion Buttons. In Proceedings of the 5th International Conference on Conversational User Interfaces (pp. 1-10).

34. Clifford Nass. 1996. Technology and Roles: A Tale of Two TVs. *Journal of Communication*, *46*(2), 121-28.
35. David Nicholas, Paul Huntington, Hamid R. Jamali, and Tom Dobrowolski. 2007. Characterising and evaluating information seeking behaviour in a digital environment: spotlight on the 'bouncer'. *Information Processing & Management*, *43*(4), 1085-1102.
36. Richard E. Petty, John T. Cacioppo, Richard E. Petty, and John T. Cacioppo. 1986. The elaboration likelihood model of persuasion. In L. Berkovitz (Ed.), Advances in experimental social psychology (Vol. 19, pp. 123–205). New York: Academic Press.
37. Robert LaRose, and Nora J. Rifon. 2007. Promoting i-saftey: Effects of privacy warnings and privacy seals on risk assessment and online privacy behavior. The Journal of Consumer Affairs, 41, pp. 127-149
38. Stephen A. Rains, and Carolyn Donnerstein Karmikel. 2009. Health information-seeking and perceptions of website credibility: Examining Web-use orientation, message characteristics, and structural features of websites. *Computers in Human Behavior*, *25*(2), 544-553.
39. Bhavika R. Ranoliya, Nidhi Raghuwanshi, and Sanjay Singh. 2017. "Chatbot for university related FAQs." In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1525-1530. IEEE.
40. Srinivasan Ratneshwar, and Shelly Chaiken. 1991. Comprehension's role in persuasion: The case of its moderating effect on the persuasive impact of source cues. *Journal of consumer research*, *18*(1), 52-62.
41. Minjin Rheu, , Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic review: Trust-building factors and implications for conversational agent design. International Journal of Human–Computer Interaction, 37(1), 81-96.
42. Soo Young Rieh. 2002. Judgment of information quality and cognitive authority in the Web. Journal of the American Society for Information Science and Technology, 53, 145–161. http://doi.org/10.1002/asi.10017
43. Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, 102551.
44. Donghee Shin. 2022. Expanding the role of trust in the experience of algorithmic journalism: User sensemaking of algorithmic heuristics in Korean users. Journalism Practice, 16(6), 1168-1191.
45. Donghee Shin. 2022. How do people judge the credibility of algorithmic sources?. *AI & Soc* 37, 81–96. https://doi.org/10.1007/s00146-021-01158-4
46. Don W. Stacks, and Michael B. Salwen (Eds.) 2014. An integrated approach to communication theory and research. Routledge.
47. Statista. 2023. 'Robo-Advisors - Worldwide'. (https://www.statista.com/outlook/dmo/fintech/digital-investment/robo-advisors/worldwide#users, accessed June 3, 2023).
48. Dirk D. Steiner, and Jeffrey S. Rain. 1989. Immediate and delayed primacy and recency effects in performance evaluation. Journal of Applied Psychology, 74(1), 136.
49. S. Shyam Sundar, Qian Xu, and Anne Oeldorf-Hirsch. 2009. Authority vs. peer: How interface cues influence users. In *CHI'09 Extended Abstracts on human factors in computing systems* (pp. 4231-4236).
50. Tze Wei Liew, Su-Mae Tan, Jessica Tee and Gerald Guan Gan Goh. 2021. "The effects of designing conversational commerce chatbots with expertise cues," 14th International Conference on Human System Interaction (HSI), Gdańsk, Poland, 2021, pp. 1-6, doi: 10.1109/HSI52170.2021.9538741.
51. Michael Tertilt, and Peter Scholz. 2017. 'To Advise, or Not to Advise — How Robo-Advisors Evaluate the Risk Preferences of Private Investors', The Journal of Wealth Management (Vol. 21), Institutional Investor Journals Umbrella, July 31.
52. Shawn Tseng, & B.J. Fogg. 1999. Credibility and computing technology. Communications of the ACM, 42(5), 39-44.
53. Julian Unkel, & Alexander Haas. 2017. The effects of credibility cues on the selection of search engine results. *Journal of the Association for Information Science and Technology*, *68*(8), 1850-1862.
54. Tavish Vaidya, Daniel Votipka, Michelle L. Mazurek, and Micah Sherr. 2019. Does being verified make you more credible? Account verification's effect on tweet credibility. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).
55. Joseph B. Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research*, *23*(1), 3-43.
56. Rebecca Walton. 2013. How trust and credibility affect technology-based development projects. Technical Communication Quarterly, 22(1), 85-102.
57. Hana Demma Wube, Sintayehu Zekarias Esubalew, Firesew Fayiso Weldesellasie, and Taye Girma Debelee. 2022. Text-based chatbot in financial sector: a systematic literature review. *Data Sci. Financ. Econ*, *2*(3), 232-259.
58. S. Xiao and Izak Benbasat. 2002. The Impact of Internalization and Familiarity on Trust and Adoption of Recommendation Agents.

unpublished Working Paper 02-MIS-006, University of British Columbia, Vancouver, Canada.
59. https://www.figma.com/
60. https://www.gov.uk/government/publications/the-national-minimum-wage-in-2023/the-national-minimum-wage-in-2023
61. https://www.ibm.com/spss/
62. https://www.prolific.co/