Towards Trustworthy AI Systems: A Human-AI Interaction Study

Completed Research Paper

Thuy-Trinh (Chloe) Nguyen
UNSW Business School
UNSW Sydney, Australia
chloe.nguyen@unsw.edu.au

Shan L. Pan
UNSW Business School
UNSW Sydney, Australia
shan.pan@unsw.edu.au

Hoang D. Nguyen
School of Computer Science and
Information Technology
University College Cork, Ireland
hn@cs.ucc.ie

Abstract

The partnership between humans and artificial intelligence (AI) has transformed decision-making and brought significant improvements in various fields. However, the complex operations of AI remain a black box in many cases, causing a lack of transparency that affects human trust in AI systems, particularly in high-risk scenarios. To address this issue, multi-agent systems have been proposed, where humans and AI interact and collaborate to achieve a better outcome and trust level. This study investigates the dynamic human-AI interaction and how it affects trust. We proposed design guidelines for interactive, trustworthy AI systems and developed two prototype versions to facilitate fake profile screening on online social networks. The study reports a mean trust score of 3.84/5 between humans and AI, despite a significant difference in their decisions on 2,142 user profiles. The results offer comprehensive insights into information systems involving human-AI interactions and underscore the increasing necessity for trustworthy AI.

Keywords: Trustworthy AI, human-AI interaction, interactive AI systems, fake identity detection, social screening

Introduction

The integration of artificial intelligence (AI) in decision-making has brought significant improvements to various aspects of life, from day-to-day human activities to corporate decision-making. In the realm of autonomous and electric vehicles, AI algorithms support the localisation of the vehicles and enhance power management and safety driving capabilities (Naz et al. 2022). In healthcare, AI systems have been studied to support healthcare professionals in dentistry for recommending tooth extraction (Khanagar et al. 2021) and in anesthesiology for enhancing intraoperative care safety (Duran et al. 2023). In corporate governance, AI is viewed as a transformative force to improve decision-making in a variety of management practices, including strategic planning, marketing, accounting and others (Gil et al. 2020). With the enormous applications of AI, it is crucial to ensure its trustworthiness.

In certain situations, human beings are susceptible to errors, whereas machines, particularly AI, possess remarkable computational capability and a vast potential to store and process voluminous and complex

data at an unparalleled speed and efficiency (Russell et al. 2017). On the other hand, while AI is useful and needed to inform decision-makers, the complex operations of AI remain a black box in most cases. The lack of transparency in AI algorithms affects the trust of humans in AI systems, especially in high-risk scenarios (Gorelik and Gyftopoulos 2020). Human intervention, hence, is necessary to oversee AI's process and handle unpredicted complex situations involving emergencies, communication, ethical dilemmas, and negotiations. It is essential to (1) recognize the complementary roles that humans and AI play in achieving optimal outcomes and (2) implement strategies that enhance their effective collaboration. This prompts the solution of multi-agent systems in which human and AI interact and collaborate to achieve a better outcome and trust level. While the phenomenon has been recognised in the literature, much must be done to take it from principles to practices (Pailian and Li 2022).

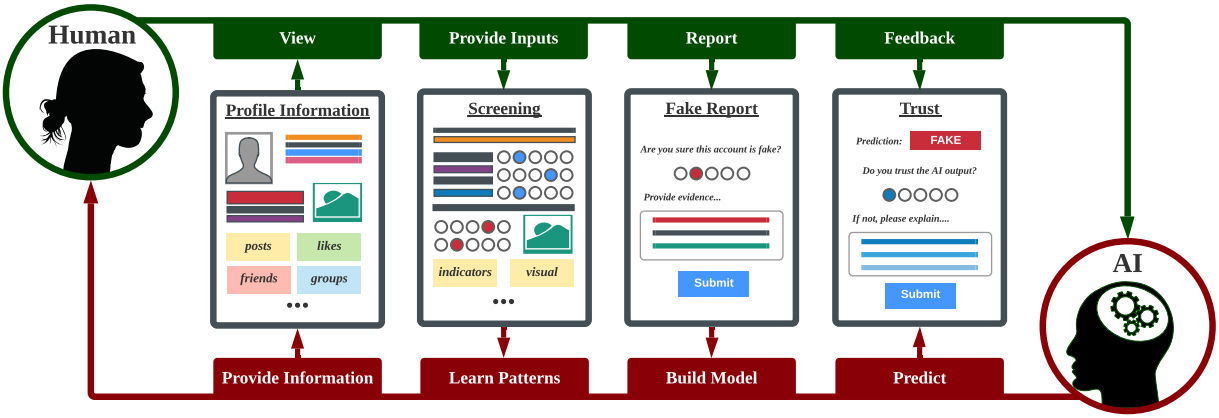

Figure 1. Study Trustworthy AI through Human-AI Interactive System for Fake Identity Detection

Online social networks (OSNs) have become the go-to platform for people to stay connected, share information, and engage with others through images, opinions, and daily interactions. However, illegitimate users can pose serious threats to social media users' privacy and security. Such users can use fake profiles, introduce malware, steal identities, and engage in various forms of harassment, which can have serious consequences (Fire et al. 2012). Identifying fake OSN profiles, however, is a challenging task that can lead to inconclusive results when conducted by humans (Shahid et al. 2022). Therefore, the use of AI can be a valuable addition to the process, given the volume of information and computational intelligence involved.

To study trustworthy AI through the impact of human-AI interaction, this study develops an interactive AI system for social screening for fake profile detection. Figure 1 illustrates the interaction workflow between humans and AI. On the first screen, our interactive AI system displays publicly available information about the profiles to human experts. Subsequently, on the second screen, human experts participate in the data annotation process, contributing insights on textual and visual indicators to the system. Afterwards, human experts are prompted to label whether the profile is fake, which is used for supervised learning. Finally, trained models are deployed to predict new data in conjunction with human judgment. Trust in AI is explicitly elicited as feedback to reinforce the human-AI synergy. It takes a human-centred approach to investigate online identities from the perspective of an ordinary user of OSNs. By incorporating humans in the process, input data can be made comprehensible to machines at the beginning and insights generated by algorithms can be double-checked and evaluated by humans in the last mile before any decision is made, which helps tackle biases and ensure trust (Silberg and Manyika 2019). Therefore, finding a synergy in which humans and AI can harmoniously work together and support each other would significantly improve the output's reliability and the system's ultimate performance. It facilitates intuitive and useful interactions between these two factors for fake profile detection in a multidimensional manner.

We examined 2,142 online identities from a social networking site between January 2020 and August 2020. The collected data contains a variety of public user information including activity logs, network information, and social relationships. On the one hand, the study employed fifteen human experts to assess the authen-

ticity of social profiles. On the other hand, our system also analysed unstructured data such as education, location information, and user-generated content. Several machine learning techniques were deployed to extract and classify fake identities based on user interactions and graph analyses of nodes, edges and categorised edges. The decisions were made on the basis of the threshold or ranking.

We introduced two versions of the interactive AI system, without and with human-AI feedback, to evaluate the potential of human-AI collaboration and human trust score. As the demand for human social intelligence is growing in Sybil detection applications due to the increasingly complicated behaviours of the attackers (Fire et al. 2012; Ramalingam and Chinnaiah 2018), we expect that humans will augment the performance of AI and improve the trustworthiness of AI.

The experimental results suggested a mean trust score of 3.84 ± 1.2 over a 5-point Likert scale between humans and AI, despite a significant difference in their decisions ($\chi^2 = 11.418, p = .001$). Also, the impact of human-AI interactions was evidently highlighted based on two versions of our interactive AI system. Our investigation with human experts indicated the mutual augmentation effect between humans and AI and the growing demand for trustworthy AI systems.

The study contributes to the understanding of trustworthy AI systems in three folds. Firstly, the paper proposes design guidelines on human-AI interaction for trustworthy AI systems. Secondly, two iterations of a prototype for social profile screening are developed to investigate the implications of the interactive system systematically. Lastly, the study offers valuable insights into the interrelationship between the confidence level of the AI model and human trust, further enhancing our understanding of trustworthy AI systems.

Related Literature

This paper investigates several related work in the literature on trustworthy AI and human-AI interaction design.

Trustworthy AI Systems

Trustworthy AI is a multifaceted concept that comprises subconcepts including AI, trust, and trustworthy AI (Kaur et al. 2022). According to Legg, Hutter, et al. (2007), AI is referred to as the ability of machines to imitate human intelligence, behaviour, and decision-making. Trust is a complex concept that has been interpreted differently across various fields. In the context of technology, trust can be defined as the confidence that the element will behave as expected (Boyens et al. 2015). Kaur et al. (2022) defines trustworthy AI as a framework to ensure that the system is trustworthy and behaves according to stakeholders' expectations.

Despite the recurring mentions of trustworthy AI in recent literature, a universally accepted set of guidelines outlining the parameters for the concept has yet to emerge. The identified properties of trustworthy AI, as highlighted in numerous sources, can be classified into two primary groups: (1) AI capability and delivery, and (2) human agency and oversight. The first group, concerning trustworthiness in AI capability, includes aspects such as accuracy, robustness, fairness, explainability, and transparency in operations (Kaur et al. 2022). These characteristics are commonly integrated into the objectives of the AI training process. Furthermore, to instill trustworthiness in AI systems, human agency and oversight are imperative. This necessitates human control in decision-making processes to mitigate unforeseeable risks effectively.

While acknowledging the positive contributions of AI to our economy and society, concerns still exist in both the short and long term. These concerns encompass but are not confined to, issues related to privacy, ethics, bias, safety, and security (Amodei et al. 2016). Noteworthy instances of discrimination and bias targeting minority populations, such as black individuals and women, have been documented in facial recognition systems (Dastin 2022) and job application screening systems (Mullen 2015). Consequently, it is strongly recommended that a human-centered interactive approach be adopted in the design of AI systems.

Recognising the shortcomings of state-of-the-art AI, we review some implications of interactive AI in different industries to understand how humans can interact with AI to overcome these obstacles. For instance, regarding the lack of transparency and explainability, with humans being involved and taking actions in an interactive AI system, it becomes harder for the process to remain hidden; thus, the transparency and

explanatory abilities issue can be tackled. This would result in smarter users and smarter learning systems as participants would gain a better understanding of the systems and how to make further improvements (Kulesza et al. 2015). As for data availability, model training, and learning capabilities, by including humans in the process, human-in-the-loop AI can help improve the overall performance of the machines as they not only learn from given training datasets but also from human expertise and feedback to emulate the knowledge-driven learning of the system for problem-solving in the future.

The roles humans would play in an interactive AI system are essential, such as controlling the input, tuning the parameters or utilising the insights generated by AI to make decisions in the last mile. Yet, another aspect that has often been neglected when discussing interactive AI systems is algorithm and system design (Boukhelifa et al. 2018). To harness the best human-AI interaction in the system, it is imperative that system design and interface facilitate intuitive and effective interactions between these two factors. Moreover, it is also suggested that design should encourage granularity and avoid building "oracles" which might provide the right answers with no explanation and little usability to end-users (Wang 2020).

Human-AI Interaction

Empowered by recent advancements in AI technologies, human-AI interaction comes in various forms, such as text suggestion, speech recognition, and computer vision technologies. Though AI-powered systems are widely implemented, reliability, transparency, and trust remain open research topics in human-AI interaction. Trust is described as the cornerstone of human-AI interaction as it determines user acceptance of AI and machine learning products. In a two-stage trust-building model between humans and AI, key AI features affecting the process are transparency in the initial trust formation phase and reliability in the continuous trust development phase (Siau and Wang 2018).

Transparency is the disclosure of the system's operation. An example of how system transparency is critical to human trust in AI is the issue of privacy. As many systems fail to express clarity in their data collection process, privacy becomes a special sub-problem of system transparency (Höök 2000), which results in the abandonment of certain products (Siau and Wang 2018) and the rise of incognito browsers and virtual private networks (Sundar 2020).

Regarding reliability, errors are expected in any AI-infused system. Albeit its impressive computational and analytical abilities to efficiently deal with the complex aspects of problem-solving, AI does not have an advantage in problems with uncertainty and equivocality issues (Jarrahi 2018). Common sense and social intelligence are beyond the capabilities of AI; therefore, it could make mistakes by suggesting actions that are against social norms. For example, an activity tracker application reminds users to stand up regardless of their social situations, such as when they are in a meeting at work or driving (Amershi et al. 2019). The issue of distributional shift discussed earlier is also at play when evaluating the reliability of AI systems. Beyond accuracy, there lies the issue of the human mental model of AI error boundaries that is *"Do people know when to accept or correct AI outputs"* (Bansal et al. 2019). If there is a mismatch between users' mental models and AI error boundaries, undetected errors in high-stake fields (e.g., autonomous traffic, healthcare) may lead to severe consequences. Failures in demonstrating transparency and reliability may lead to refusal of the AI products (De Graaf et al. 2017). Subsequently, modern users demand a comprehensive human-AI interaction design where transparency and reliability in AI-powered systems are prominent features.

In the last 20 years, a growing body of research has discussed human-AI interaction design. The pioneering attempts outline classic concerns, including the sense of human control (e.g., transparency, privacy, predictability) and output verification (Höök 2000; Norman 1994). One recent work by Amershi et al. (2019), whose authors were inspired by principles of mixed-initiative user interfaces (Horvitz 1999), provides a more complete reusable set of human-AI interaction guidelines. This work codifies scattered recommendations from previous papers and provides 18 principles reviewed on real-life AI products ranging from activity trackers, email plugins, and navigation applications. These guidelines are categorised into four general stages: initial setup (e.g., how to set the right expectations), during interaction (e.g., how to handle services based on social context), when AI is wrong (e.g., how to include feedback from users), and overtime interactions (e.g., how the AI system learns from users) (Amershi et al. 2019).

One crucial feature shared by various human-AI interaction design frameworks is the idea of incorporating

user feedback into AI-infused systems, which improves not only the system's transparency but also its reliability. Besides the concept of intelligence amplification that AI can augment human abilities (Rheingold 2000), human input is also valuable to AI-driven insights. As described earlier, AI systems are prone to producing suboptimal outputs in situations involving high uncertainty or social intelligence. Therefore, by building a validation step to safeguard the output quality, users are more aware of how AI makes decisions and are able to accept or override its recommendations to ensure the system's accuracy. The performance of human-AI synergy potentially excels their individual efforts as it overcomes the limitations of both agencies (Sundar 2020).

Methodology

This section discusses our research objectives for designing a trustworthy system with human-AI interactions for fake identity detection. The section explains our experimental design, data collection and analysis.

Research Objectives

The study aims to investigate the impact of interactions between humans and AI towards trustworthy AI in a real-world application for identifying fake profiles on OSNs. We determine trust as the cornerstone of such interactions to understand user perception of the system and AI outputs. Hence, we examine the following prominent research questions:

RQ1. Is there a good level of agreement and trust between humans and AI?

RQ2. What is the impact of human-AI interactions on fake identity detection?

To explore our research questions, we developed an interactive AI system for social profile screening, promoting collaboration between human experts and AI within a user-friendly web interface. Our study involved 2,142 online identities from a popular OSN, utilising machine learning techniques like Support Vector Machines (SVMs) and neural networks in a hybrid model to identify illegitimate users. We introduced two system versions in our experiment—one with human-AI interaction and one without—to assess the impact on human trust scores.

Interactive AI System for Social Screening

OSNs profoundly impact how people interact, stay in touch, and exchange information about their lives. Given the enormous volume of personal information exchanged between friends on OSNs, protecting individual users from exploiting personal identity and data has attracted broad research attention. Previous studies have shown how friend network information can be misused to launch two types of identity theft attacks that are difficult to prevent: identity cloning attack (ICA) and fake profile attack (FPA) (Bilge et al. 2009). Both cloning strategies use the personal details of OSN existing users (e.g. name, photos, education information) to create one or more clone accounts claiming the same identity as those of the benign users. The main difference between ICA and FPA is that while the former is a single-site attack (i.e. attackers operate on the same site as the authentic users), the latter is a cross-site attack (i.e. cloning attacks happen on different sites). Besides cloning attacks, fake profiles can be developed using fabricated identities (Krombholz et al. 2012). The three types of state-of-the-art fake profile detection approaches are graph-based (network structure), content-based (feature-based), and hybrid models.

Graph-based approach. Introduced in pioneering studies of Sybil detection, network structure remains the dominant approach among the three techniques. Several successful methods that deploy and develop this approach include Gatekeeper (Tran et al. 2011) and SybilRank (Cao et al. 2012). These models focus on the links (e.g., relationships, friend lists) and nodes of the profiles to deploy algorithms such as page rank and random walk to rank OSN users. The majority of graph-based models assume that benign users are tightly connected and fake accounts have few attack edges (i.e., fake accounts can only befriend a few legitimate users). This assumption, however, has been considered oversimplified for real-life networks (Yang et al. 2014).

Content-based approach. Feature-based fake profile detection methods extract content information

(e.g. user posts, news feeds) and apply machine learning algorithms, Support Vector Machines (SVMs) and other clustering techniques to classify and rank users. Examples of this approach are TrustBook (Noor et al. 2013) and Empirical evaluation model (Ji et al. 2016).

Hybrid approach. The hybrid approach has been introduced in recent years via the applications of Sybil-Frame (Gao et al. 2015) and Íntegro (Boshmaf et al. 2015), which focus on both relationship and feature information of the users. Although the field is full of potential, research on the hybrid model is still in its infancy.

Despite the growing body of Sybil accounts detection research and its achievements, several problems arise with the expansion of OSNs.

Big data challenges. As the number of OSNs and their users is increasing impressively fast, big data becomes the new challenge for fake profile detection. Aside from classic success factors, the Sybil detection methods in this era must enhance their data processing capability, reduce computational cost and consider implementing big data computing techniques for batch processing (Ramalingam and Chinnaiah 2018). The growth of the networks also intensifies the importance of the next issue, which is the lack of preventive detection methods.

Lack of preventive methods. Among the discussed techniques, TrustBook (Noor et al. 2013) is one of the few models that focuses on preventing the formation of fake profiles with the suggestion of digital trust certificates. While the others solely tackle these Sybil accounts by analysing and detecting them, which could be more costly than resolving the problem in the first place.

Reliability and human interference. Prior studies have shown that fake profiles are becoming more challenging to detect as attackers invest more effort in mimicking benign users. For example, ordinary Facebook users have difficulties detecting fake profiles (Krombholz et al. 2012). With their connected friend networks and active behaviours, modern fake profiles impose great difficulties upon both legacy graph-based and feature-based techniques. Hence, the idea of involving trained human experts in the process of fake profile detection is highly in demand.

System Design

In this study, we design an interactive AI system that provides a mobile-friendly web interface for human-AI interactions to investigate OSN profiles. The system bundles comprehensive profile information for analysis, structured questionnaires for profile screening based on four parts (i.e., work, education, social behaviour, interest), and fake identity reports based on 5-point Likert scales. If the profile is deemed fake, human experts will explain why they believe the profile is fake. Then, the system will prompt the human experts to express their trust level in the machine learning results. We implement an average voting method to ensure that the result is accurate based on the evaluation of at least three human annotations per profile.

First, human experts evaluate the profiles. These labelled data are then fed through the AI system as training inputs, allowing AI to learn and predict users' authenticity. When AI generates its output, the results are validated by our human workers, who will rate their trust score and provide feedback for AI based on their social experience and knowledge. Finally, human feedback is incorporated into the AI system to reinforce its learning. This process represents the interdependent relationship between humans and machines in human-in-the-loop AI, which is illustrated in Figure 2.

As a building block of any comprehensive interactive AI system, we build several AI models to detect OSN fake profiles. The fundamental information of each profile, including demographics, interests, photos, posts, and joined groups, is collected through the APIs of the target OSN. The data is later standardised by removing outliers, fixing typos and grouping values. Subsequently, it is fed through a feature generator to generate 45 features as inputs for the AI models. Table 1 summarises the list of features for the AI model.

Combined with the labels (i.e., *fake* or *not fake*) annotated manually in the annotation phase, we investigate and compare several AI models, including SVM with linear kernel and multi-layer perceptron for fake identity detection. Given an arbitrary profile, each model will predict a real-value score, which represents the profile's probability of being fake. Empirically, we apply a confident threshold to classify fake profiles.

Group	Features
General	'has_name', 'has_address', 'has_location', 'has_hometown', 'has_age_range', 'has_birthday', 'has_phone', 'gender', 'relationship_status'
Work	'nb_work'
Education	'nb_education'
Social Behaviour	'nb_friends', 'nb_posts', 'nb_posts_with_tags', 'nb_mobile_photos', 'nb_profile_photos', 'nb_cover_photos', 'nb_wall_photos', 'nb_app_photos', 'nb_albums', 'nb_normal_albums', 'nb_total_photos', 'nb_event_post', 'nb_note_post', 'nb_album_post', 'nb_normal_photos', 'nb_lightweight_album_photos', 'nb_comments', 'nb_post_likes', 'nb_photo_post', 'nb_video_post', 'nb_status_post', 'nb_link_post'
Interests	'nb_liked_pages', 'nb_tagged_place', 'nb_likes', 'nb_music', 'nb_sports', 'nb_inspirational_people', 'nb_joined_groups', 'nb_closed_groups', 'nb_open_groups', 'nb_admin_groups', 'nb_books', 'nb_favorite_athletes', 'nb_favorite_teams', 'nb_games',
Label	'label'

Table 1. List of Features

Stage	Guideline	Application in Interactive System
Initially	Make clear what the system can do	Initial training and briefing sessions with human experts ensure that the system functionalities are communicated clearly.
During interaction	Show contextually relevant information	Two types of input confirmation screen (Screen 2) are displayed based on the result from the annotation screen (Screen 1) If the human expert reports a profile to be fake, Screen 2 asks for the human expert's decision confidence level and their reasonings. If the human expert does not report a profile to be fake, Screen 2 asks for confirmation of the inputs.
When AI is wrong	Support efficient invocation Make clear why the system did what it did	The instant "Report fake profile" button is designed in red and displayed at the top of Screen 1. Therefore, the human expert can report fake profiles in one click. The annotation screen (Screen 1) displays the criteria that will be used to predict fake accounts. By validating the profiles using the same criteria, human experts are aware of AI's decision basis.
Overtime	Encourage granular feedback Learn from user behaviours	Human-machine agreement screen (Screen 4) is designed to capture users' feedback. When AI displays its prediction on the genuineness of the account, the human expert can rate their level of agreement with AI output on a 5-point Likert scale. A text box is included in Screen 4 for users to input their feedback. Users' feedback from Screen 4 (both in form of Likert scale values and text) is incorporated into the system to enhance its learning.

Table 2. Design Guidelines

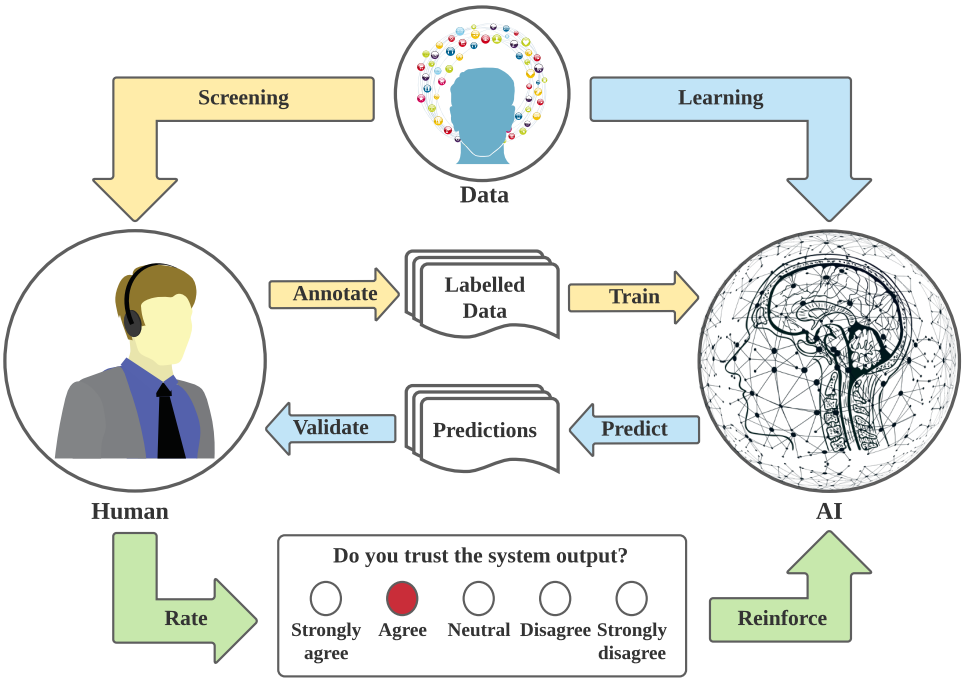

Figure 2. Human-AI Interactions for Social Screening

In social screening, the use of machine learning provides an efficient way to analyse the enormous amount of profile data. In contrast, human experts can address multifaceted issues, such as implicit social cues, complicated relationship statuses, or forged backgrounds. Therefore, the system is designed to best facilitate interactions between AI and humans with millions of digital footprints, including general information (e.g., work, education, phone, gender, address, and location), photos, group activities, likes, interests, and posts. A digital identity reassembles all the characteristics and social behaviours of a person, enabling both humans and AI to investigate fake accounts.

We follow the human-AI interaction design guidelines from Amershi et al. (2019) to build the interactive AI-enabled system. As these guidelines are optimised for generality, the focus is to tailor some to be more application-specific to our fake account detection system. Table 2 demonstrates our adaptation of the guidelines. The following highlights the key design features of the system.

Profile Information and Screening. This module is structured to provide all the information on the profile and screening questions. The profile section on the left-hand side displays public information, including general information, images, and user posts. Located on the right-hand side is a series of both fixed-alternative and open-ended questions to evaluate the profiles, as illustrated in Figure 3. These fixed alternatives guide human experts in investigating digital footprints and understanding profile authenticity.

Fake Profile Report. In this step, the human experts are prompted to identify if the profile is fake and their confidence using a 5-point Likert scale. A text box is provided for the experts to explain their reasoning, as illustrated in Figure 4.

AI Prediction and Trust. In the machine agreement step, the system displays its prediction on whether the account is fake, as shown in Figure 5. The human experts are required to rate their degree of trust in the output of AI using a 5-point Likert scale (Strongly agree - Agree - Neutral - Disagree - Strongly disagree) and enter their comments in a text box. The user's feedback can override AI recommendations and will be incorporated into the system to improve its accuracy.

We launched two versions of the system. The first version (V1) showcases two main features, which correspond to Profile Screening and Fake Profile Report screens. While the former screen allows human experts

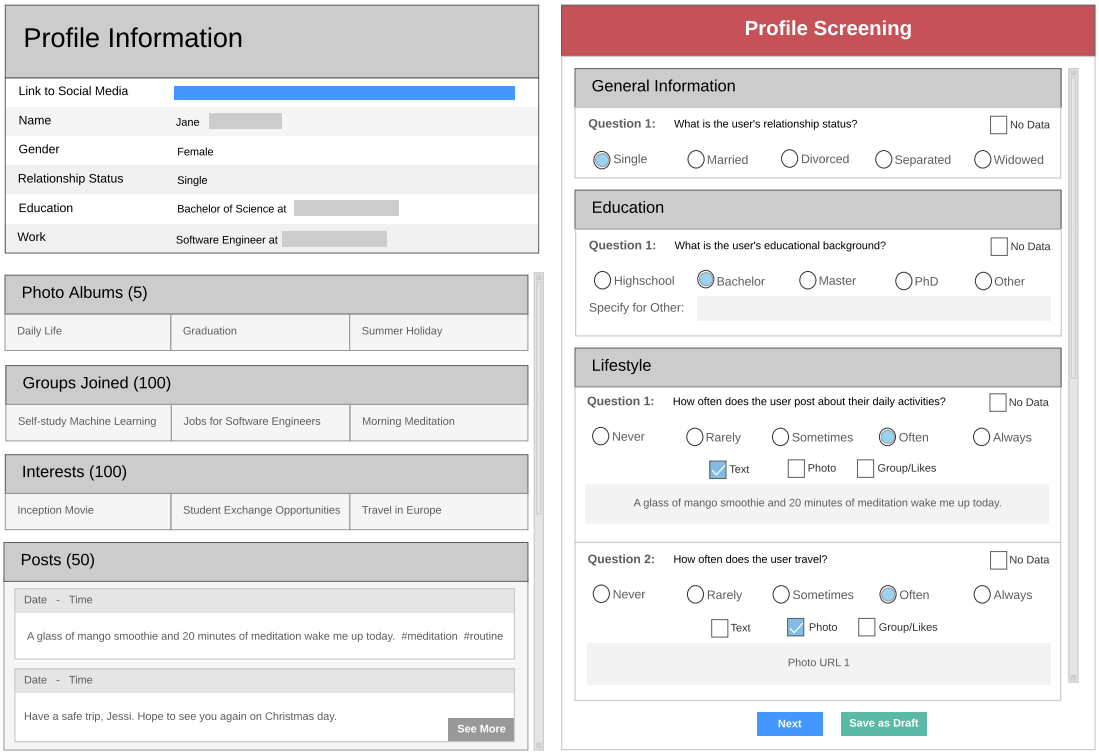

Figure 3. Profile Information and Screening (Left: Screen 1, Right: Screen 2)

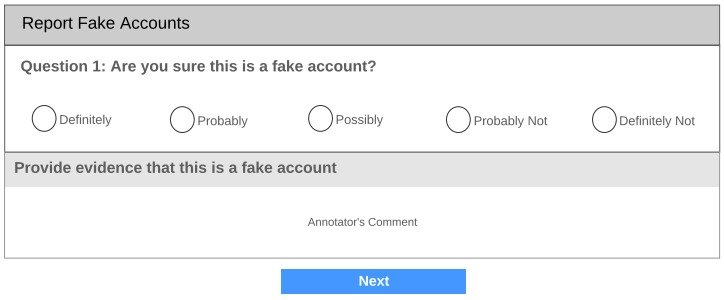

Figure 4. Fake Profile Report (Screen 3)

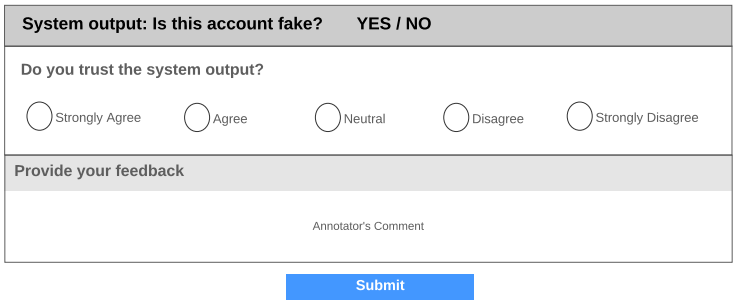

Figure 5. AI Prediction and Trust (Screen 4)

to highlight key information about the profiles, the latter provides a convenient shortcut to report highly probable fake profiles immediately. V1 of the system reflects human judgements without any external suggestions. After three weeks into the fake profile annotation process, we introduced the second version of the system (V2). V2 has the full features of V1, but with the addition of the third module, a human-AI agreement screen. As the last step of the annotation process, the system displays the authenticity prediction of the profiles to human experts. The potential impact of V2 is that it allows human experts to provide feedback to the system and, most importantly, rate their trust score. We are also interested in learning whether humans and AI can augment each other's intelligence in the application of fake profile detection. In a favourable scenario, we expect that the additional feature in V2 will enhance both AI by incorporating human social and online experiences through training and human performance.

Experimental Design

Our experiments involve human experts and the AI algorithm evaluating the user profiles based on comprehensive profile data, including profile information, activities, friends, groups, posts, and images. The system provided human experts with profile data and a list of structured questions to guide them in screening OSN identities. It is important for human experts to collect evidence and develop valid reasoning for decisions. With the goal of increasing granularity in the feedback process, we wanted to use a system with more intricate outputs than binary classification alone. Therefore, we designed the questions on 5-point Likert scales. The task was to evaluate the authenticity of over 2000 profiles based on their information. It is the standard protocol that each profile be annotated by at least three human experts. Then, the annotation results will be finalised using average voting. In the next steps, human experts will be prompted to detect whether the profile should be marked as fake and to provide reasons for their decision. Finally, there are two versions of the system: V1 - without human-AI interaction, and V2 - with human-AI interaction.

AI results will be disclosed to human experts in human-AI interactions to elicit their trust in the machine's answer. Because our main metric (i.e., the perception of model accuracy and user trust) is based on experts' experience with the system, we decided that each participant should not see the AI results before reaching their decision as it may deviate from their answers. The first independent variable in our experiment is the interaction presence at two levels: *with-interaction* and *without-interaction*. Subjects in the *with-interaction* condition were asked to rate their trust score for model results and the reasons for their choice. Participants in *with-interaction* condition were explicitly instructed that their feedback would be used by the model to update the model's parameters and reinforce machine learning.

Data Collection

The study was conducted from January to August 2020. The V1 of the system was deployed during the first three weeks. We gathered the results from a total of 1,576 unique profiles and 4,405 annotation times by fifteen human experts. Data collection was performed through random crawling and human annotators with diverse backgrounds were recruited, both measures taken to minimise bias. During the last two weeks, the same group of human experts worked on the V2 of the system with the facilitation of human-AI interaction. There were 440 unique profiles and 1,693 annotations done in conjunction with AI predictions. In total, we collected 2,016 unique profiles out of 2,142 profiles with 6,098 annotation times on 45 questions.

Data Analysis

Before conducting any statistical analysis, some key data fields (e.g., human trust scores) were normalised. Our main purposes are to (1) compare the annotation results of human experts and AI, (2) compare human experts' performance in the two versions (with and without human-AI interaction), and (3) evaluate human trust score. In order to satisfy these aims and reveal insights into the data, we focused on tests of difference, including chi-square and Kolmogorov–Smirnov (KS) tests.

Gender	**n**	**%**	**Marital status**	**n**	**%**
Male	1167	57.8%	Single	757	42.1%
Female	846	41.9%	Married	922	51.3%
Others	7	0.3%	Divorced	16	0.9%
			Others	103	5.7%
Number of dependants	**n**	**%**	**Living condition**	**n**	**%**
1-2	668	63.1%	Low	79	4.4%
3-4	246	23.2%	Medium	943	52.5%
Others	145	13.7%	High	697	38.8%
Profession	**n**	**%**	**Education**	**n**	**%**
Freelancer	449	29.4%	High school	469	27.6%
White-collar	888	58.2%	Bachelor	875	51.4%
Middle-level Manager	20	1.3%	Master degree	16	0.9%
High-level Manager	22	1.4%	Others	341	20.1%
Others	148	9.7%			
Social status	**n**	**%**	**Own an online business**	**n**	**%**
Do not have	1716	85.3%	Do not have	853	41.3%
Do have	55	2.7%	Do have	1097	53.1%

Table 3. Profile Demographics

Results and Discussions

In this section, we present the measures of our user studies using quantitative method. Statistical test results are reported based on chi-square and KS tests to answer our research questions.

Results

We conducted a number of quantitative analyses for both versions of the system, V1 (without Human-AI feedback) and V2 (with Human-AI feedback), to address **RQ1** and **RQ2**. Our aims are to (1) assess which demographic factors will affect the fake detection decision, (2) compare the difference between human experts and machine results, (3) conduct an analysis of human trust in machine results, (4) assess the effect of human and AI feedback screen on the decision of human experts (5) and make the comparison between AI confidence level and trust score.

Demographics

Table 3 demonstrates that the percentage of male users is slightly greater than its counterpart and 51.3% of them have been married. It can also be seen from the demographic table that 63.1% of the profiles have 1-2 dependants while the number of those living in larger families with more than 3 members is much lower, with approximately 23.2% of the total number of profiles. The predominant profession in the sample is white-collar workers (58.2%), followed by people working as freelancers (29.4%), and the majority of them do not have any significant status in society. As profiles have disposable income from their professions and businesses, it is more likely that they can afford to be surrounded by a better living environment, with 52.5% of them having a medium living standard and the percentage of people living in a high standard 38.8%.

Our assessment results revealed that some demographic factors significantly correlated to the likelihood of fake identity. For instance, gender, social status, professional status, number of dependents, and living conditions play a vital role in predicting whether the account is fake ($p < 0.05$). We reported the adjusted R^2 of 0.296 hinting that 29.6% of the variance in the fake indication can be predicted from the demographic variables.

Comparison between human experts and machine annotations

In our analysis, human experts annotated 217 fake and 1,738 genuine profiles, while our machine learning algorithms classified 413 fake and 1,542 real users. Due to random crawling, we removed profiles that were duplicated and blank profiles that did not have sufficient public data for annotation. The reported profiles for analysis are unique and completed profiles. The confusion matrix between human and machine annotations is depicted in Figure 6. The computed F1 score of 0.21, recall of 0.3, and precision of 0.16 were relatively low, indicating a noticeable degree of discordance. Furthermore, the chi-square test confirmed a significant difference in opinions between human experts and AI ($\chi^2 = 11.418, p = .001 < 0.05$).

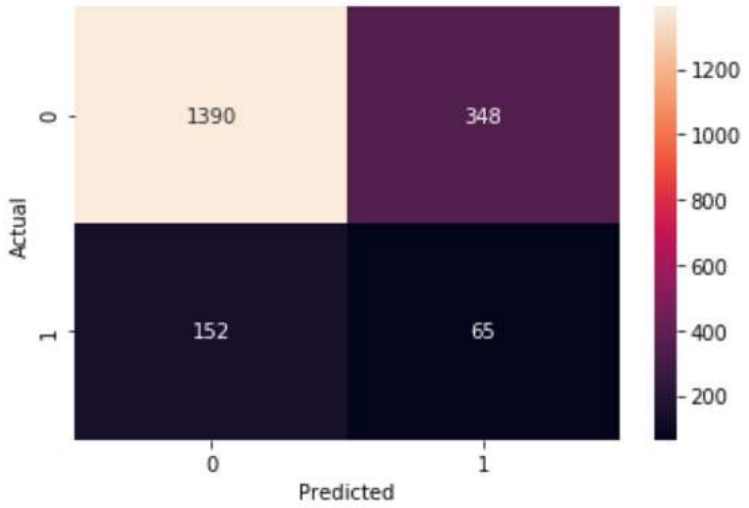

Figure 6. Analysis of Human and AI Annotation Results

Analysis of human trust on AI outputs

Despite the difference in fake profile detection, we found that the average trust score of Human AI feedback is 3.84 (SD=1.23) based on the 5-point Likert scale, as shown in Figure 7. These results indicate a noteworthy trend wherein human experts typically concur with AI outputs presented in the feedback form.

Impact of human and AI feedback screen on the decision of human experts

In this analysis, we compared human annotations between two versions, V1 and V2, to understand the impact of AI predictions on human expert decisions. A chi-square test was performed for two versions of the system. The p-value of our test was smaller than 0.05, which strongly indicates that the human annotation results with and without human-AI feedback are significantly different ($\chi^2 = 2016.0, p = .000$). This suggests that after going through the Human AI Feedback screen, human opinions would be deviated and affected by the machine results.

Given that the chi-square test can accommodate variations in sample sizes across groups without compromising its validity, and considering that both versions contain sufficiently large samples that meet this assumption, it can be concluded that the differences in sample sizes do not impact the test's results.

The relationship between AI confidence level and human trust score

In order to ascertain the correlation between the AI confidence level and human trust score in relation to AI, a Kolmogorov–Smirnov (KS) test was conducted on the AI confidence level and the normalised trust score, as depicted in Figure 8. The outcomes of the KS test indicate a significant disparity between the trust score and AI confidence level, signifying that humans tend to place trust in AI results even if the confidence level is low.

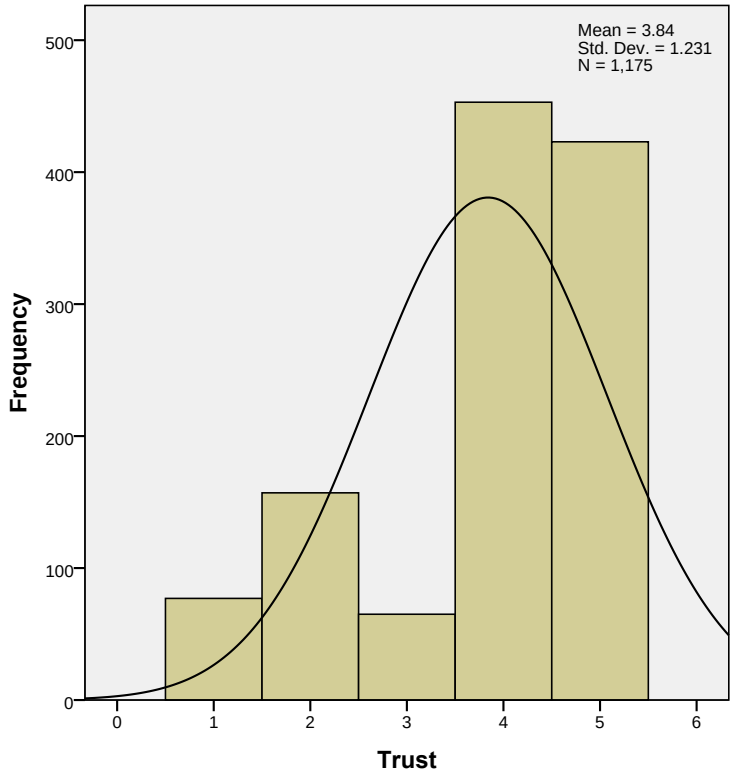

Figure 7. Trust on AI

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Normal machine agreement	Statistic	df	Sig.	Statistic	df	Sig.
Normal machine confidence	.00	.156	77	.000	.812	77	.000
	.25	.170	157	.000	.759	157	.000
	.50	.261	65	.000	.580	65	.000
	.75	.156	453	.000	.694	453	.000
	1.00	.180	423	.000	.684	423	.000

a. Lilliefors Significance Correction

Figure 8. Assessment Correlation Between AI Confidence Level and Trust Score

Discussion

This section discusses the noteworthy emerging topics from the experiment insights and the study's implications for trustworthy AI research enthusiasts and the broader information systems research community.

Multifaceted Aspects of Trustworthy AI

Despite differences between human and AI results, users tend to trust AI outcomes. This inclination can be explained by two main factors. Firstly, users may lack full confidence in their own assessments, considering that AI incorporates the opinions of two additional human experts for each profile annotation. This fosters the perception that AI possesses superior skills in detecting fake profiles. Secondly, there is a prevailing belief that machines are adept at recognising micro-features, surpassing human capabilities in certain aspects. Consequently, users extend their trust to AI-generated answers. The concept of trust is multifaceted; therefore, AI trustworthiness encompasses not only accuracy and human supervision, which are the focus

of this study, but also explainability and transparency, among other factors.

AI confidence and human trust are not the same. The AI confidence score denotes the system-generated probability associated with each predicted label, reflecting the algorithm's confidence in its output. On the other hand, the human trust score represents the level of confidence humans have in the AI-generated output. Therefore, these two metrics can diverge in many instances. This discrepancy arises due to the fact that the results of the AI confidence score are not disclosed to humans, leading them to perceive AI annotation results as superior and, subsequently, place trust in AI answers. Most studies in the literature, including this research, adopt black-box AI algorithms, which do not adequately disclose their internal mechanisms. This underscores the necessity for greater transparency in the operations of AI to enhance human decision-making processes.

Human-AI Interaction and Knowledge Augmentation

Human-AI interaction makes a significant difference. The interaction between humans and AI through feedback mechanisms facilitates users' critical evaluation of their decisions, prompting a reconsideration of their responses in light of the rationales underlying AI-generated results. Consequently, users may iteratively adjust their answers following a validation process with AI outcomes. This dynamic human-AI interaction serves as a catalyst for both human users and the AI system to fortify their knowledge base. Further considerations for human-AI interaction design include multi-level interactions (e.g., multiple rounds of feedback and decision-making refinements) and multi-stage interactions (e.g., collaboration in data input and evaluation).

The integration of human and AI intelligence represents a promising avenue for knowledge augmentation. By leveraging the strengths of each, such hybrid systems have the potential to yield more effective outcomes and facilitate better decision-making. Human social intelligence and intuition play a crucial role in the identification of fake profiles, particularly in the context of unstructured data, such as social data. The introduction of a human-AI feedback interface prompts individuals to comprehend the rationale behind machine-generated results, potentially augmenting the precision of annotations. This reciprocal interaction encourages humans to reconsider and enhance their responses, contributing to knowledge enhancement. Additionally, AI can leverage human social intelligence to enhance its accuracy.

Implications

The study presents a set of design principles tailored for the development of trustworthy AI systems empowered by human-AI interactions. These principles provide information systems practitioners with practical guidelines for constructing reliable AI solutions. The validated results indicate that the use of the design elements enhances trust. This paper lays the groundwork for future research, providing a stepping stone for further exploration into collaborative human-AI frameworks.

The applicability of this research extends beyond social screening, where relying solely on human annotators often yields inconclusive results (Krombholz et al. 2012). Researchers in the information systems field can utilise the study's insights to address scenarios where augmented intelligence is needed, providing assistance in areas like fraud detection, forecasting, and complex multidimensional decision-making processes.

Conclusion

Trustworthy AI is an emerging paradigm that is gradually becoming an indispensable part of AI applications. As such, it is crucial to pay attention to the partnership and interaction between humans and AI. This study provides design principles for trustworthy AI systems and develops an interactive system for social screening through fake profile detection. The experimental results highlight that trust played a critical role despite AI's certain shortcomings. The study demonstrates the possibility of augmented intelligence, where humans and AI can leverage each other's strengths to achieve better outcomes overall. This work helps bridge the gap between principles and practices in implementing interactive human-AI factors to enhance AI trustworthiness. The study provides essential insights for researchers aiming to navigate the evolving landscape of trustworthy AI in information systems.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. (2019). "Guidelines for human-AI interaction," in *Proceedings of the 2019 chi conference on human factors in computing systems,* pp. 1–13.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565* ().

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019). "Beyond accuracy: The role of mental models in human-AI team performance," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing,* vol. 7. 1, pp. 2–11.

Bilge, L., Strufe, T., Balzarotti, D., and Kirda, E. (2009). "All your contacts are belong to us: automated identity theft attacks on social networks," in *Proceedings of the 18th international conference on World wide web,* pp. 551–560.

Boshmaf, Y., Logothetis, D., Siganos, G., Lería, J., Lorenzo, J., Ripeanu, M., and Beznosov, K. (2015). "Integro: leveraging victim prediction for robust fake account detection in OSNs." in *Ndss,* vol. 15. Citeseer, pp. 8–11.

Boukhelifa, N., Bezerianos, A., and Lutton, E. (2018). "Evaluation of interactive machine learning systems," in *Human and Machine Learning,* Springer, pp. 341–360.

Boyens, J., Paulsen, C., Moorthy, R., Bartol, N., and Shankles, S. A. (2015). "Supply chain risk management practices for federal information systems and organizations," *NIST Special publication* (800:161), p. 32.

Cao, Q., Sirivianos, M., Yang, X., and Pregueiro, T. (2012). "Aiding the detection of fake accounts in large scale social online services," in *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12),* pp. 197–210.

Dastin, J. (2022). "Amazon scraps secret AI recruiting tool that showed bias against women," in *Ethics of data and analytics,* Auerbach Publications, pp. 296–299.

De Graaf, M., Allouch, S. B., and Van Diik, J. (2017). "Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI,* IEEE, pp. 224–233.

Duran, H., Kingeter, M., Reale, C., Weinger, M. B., and Salwei, M. E. (2023). "Decision-making in anesthesiology: will artificial intelligence make intraoperative care safer?," *Current Opinion in Anaesthesiology* (36) 6 2023, pp. 691–697.

Fire, M., Katz, G., and Elovici, Y. (2012). "Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies," *Human Journal* (1:1), pp. 26–39.

Gao, P., Gong, N. Z., Kulkarni, S., Thomas, K., and Mittal, P. (2015). "Sybilframe: A defense-in-depth framework for structure-based sybil detection," *arXiv preprint arXiv:1503.02985* ().

Gil, D., Hobson, S., Mojsilović, A., Puri, R., and Smith, J. R. (2020). "AI for management: An overview," *The future of management in an AI world: Redefining purpose and strategy in the fourth industrial revolution* (), pp. 3–19.

Gorelik, N. and Gyftopoulos, S. (2020). "Applications of artificial intelligence in musculoskeletal imaging: from the request to the report," *Canadian Association of Radiologists Journal* (72) 1 2020, pp. 45–59.

Höök, K. (2000). "Steps to take before intelligent user interfaces become real," *Interacting with computers* (12:4), pp. 409–426.

Horvitz, E. (1999). "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems,* pp. 159–166.

Jarrahi, M. H. (2018). "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Business Horizons* (61:4), pp. 577–586.

Ji, Y., He, Y., Jiang, X., Cao, J., and Li, Q. (2016). "Combating the evasion mechanisms of social bots," *computers & security* (58), pp. 230–249.

Kaur, D., Uslu, S., Rittichier, K. J., and Durresi, A. (2022). "Trustworthy artificial intelligence: a review," *ACM Computing Surveys (CSUR)* (55:2), pp. 1–38.

Khanagar, S., Al-Ehaideb, A., Maganur, P. C., Vishwanathaiah, S., Patil, S., Baeshen, H. A., Sarode, S. C., and Bhandi, S. (2021). "Developments, application, and performance of artificial intelligence in dentistry – a systematic review," *Journal of Dental Sciences* (16) 1 2021, pp. 508–522.

Krombholz, K., Merkl, D., and Weippl, E. (2012). "Fake identities in social media: A case study on the sustainability of the Facebook business model," *Journal of Service Science Research* (4:2), pp. 175–212.

Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. (2015). "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th international conference on intelligent user interfaces,* pp. 126–137.

Legg, S., Hutter, M., et al. (2007). "A collection of definitions of intelligence," *Frontiers in Artificial Intelligence and applications* (157), p. 17.

Mullen, J. (2015). *Google rushes to fix software that tagged photo with racial slur. CNN*.

Naz, N., Ehsan, M. K., Amirzada, M. R., Ali, M. Y., and Qureshi, M. A. (2022). "Intelligence of autonomous vehicles: a concise revisit," *Journal of Sensors* (2022), pp. 1–11.

Noor, U., Anwar, Z., Mehmood, Y., and Aslam, W. (2013). "TrustBook: web of trust based relationship establishment in online social networks," in *2013 11th International Conference on Frontiers of Information Technology,* IEEE, pp. 223–228.

Norman, D. A. (1994). "How might people interact with agents," *Communications of the ACM* (37:7), pp. 68–71.

Pailian, H. and Li, L. (2022). "Landscape of user-centered design practices for fostering trustworthy human-ai interactions," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (66) 1 2022, pp. 1255–1259.

Ramalingam, D. and Chinnaiah, V. (2018). "Fake profile detection techniques in large-scale online social networks: A comprehensive review," *Computers & Electrical Engineering* (65), pp. 165–177.

Rheingold, H. (2000). *Tools for thought: The history and future of mind-expanding technology,* MIT Press.

Russell, S., Moskowitz, I. S., and Raglin, A. (2017). "Human information interaction, artificial intelligence, and errors," *Autonomy and Artificial Intelligence: A Threat or Savior?* (), pp. 71–101.

Shahid, W., Li, Y., Staples, D., Amin, G., Hakak, S., and Ghorbani, A. A. (2022). "Are you a cyborg, bot or human?—a survey on detecting fake news spreaders," *IEEE Access* (10), pp. 27069–27083.

Siau, K. and Wang, W. (2018). "Building trust in artificial intelligence, machine learning, and robotics," *Cutter Business Technology Journal* (31:2), pp. 47–53.

Silberg, J. and Manyika, J. (2019). "Notes from the AI frontier: Tackling bias in AI (and in humans)," *McKinsey Global Institute (June 2019)* ().

Sundar, S. S. (2020). "Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII)," *Journal of Computer-Mediated Communication* (25:1), pp. 74–88.

Tran, N., Li, J., Subramanian, L., and Chow, S. S. (2011). "Optimal sybil-resilient node admission control," in *2011 Proceedings IEEE INFOCOM,* IEEE, pp. 3218–3226.

Wang, G. (2020). *Humans in the Loop: The Design of Interactive AI Systems*.

Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., and Dai, Y. (2014). "Uncovering social network sybils in the wild," *ACM Transactions on Knowledge Discovery from Data (TKDD)* (8:1), pp. 1–29.