Can Humans Detect the Authenticity of Social Media Accounts?

On the impact of verbal and non-verbal cues on credibility judgements of Twitter profiles

Christopher Sandy
School of Psychology
University of Surrey
Guildford, Surrey, UK
chrissandy1234@hotmail.com

Patrice Rusconi
School of Psychology
University of Surrey
Guildford, Surrey, UK
p.rusconi@surrey.ac.uk

Shujun Li
Surrey Centre for Cyber Security
University of Surrey
Guildford, Surrey, UK
shujun.li@surrey.ac.uk

***Abstract*—This study investigates the influence of verbal and non-verbal cues on people's credibility judgments of fake Twitter profiles generated by an information hiding mobile app solely for transmitting secret messages. We tested the hypotheses that the trustworthiness conveyed by the profile picture, morality-related trait adjectives included in the profile summary and the profile owner's gender would increase people's credibility judgments of those fake Twitter profiles. 24 participants assessed 16 fake profiles on their credibility. They also expressed their confidence in their credibility judgements and they answered an open-ended question on which parts of the profile influenced their credibility judgements. The results showed that overall participants did not trust the Twitter profiles. Furthermore, confidence judgements were higher when profiles included competence-related traits in the profile summaries. Verbal rather than non-verbal cues had thus more influence on participants' judgements. The open-ended responses revealed a large reliance on the content of the profile, which is what the mobile app relies on. We discussed these findings in light of the relative lack of credibility of the profiles generated by the mobile app. The new insights can help improve designs of systems depending on automated social media accounts and will provide useful clues about other applications where cognitive computing plays a role.**

Keywords—Twitter; social media; twitterbots; information hiding; mobile computing; credibility; confidence; morality; competence; cognitive computing.

I. INTRODUCTION

Fake profiles and social media robots are widespread and they have considerable repercussions for individuals [1]. These include cyberbullying [2], harassment, distribution of viruses, and extraction of sensitive information [3].

Most people believe that online deception is widespread [4]. Social media accounts that automatically produce content (bots) are increasing as shown in the debate for the UK referendum on EU membership [5]. Furthermore, recent research has shown that twitterbots can be judged as credible as human agents [6]. However, there is still need of systematic empirical investigations of what factors determine human judgements of credibility. Research on the psychological aspects of authenticity judgements bears implications for the fields of cyber security and human-computer interaction.

In a bid to create a crowdsourced way of detecting fake profiles, Wang *et al.* [7] presented people with numerous fake social media profiles. They asked them if the profile was real or fake, and which part of the profile they used to arrive at their conclusion. They found that even non-expert computer users were very good at discerning the real from the fake. There is some research in the cyber security domain attempting to assess what makes fake profiles more or less deceptive. People use a lot of cues to gauge trustworthiness of profiles. These include the types and number of "friends" (social profile connections), the profile summary and people smiling in their profile pictures [8]. Furthermore, research has shown that the perceived credibility of a Twitter page's owner is lower when there are too few or too many followers [9]. Source credibility is also affected by the recency of updates on a Twitter page: faster updates trigger the user's cognitive elaboration that induces increased credibility [10]. Knowing what makes profiles deceptive helps preventing succumbing to fake profiles and enhancing the authenticity of pro-social fake profiles.

Li and Ho [11] have developed a new information hiding technology prototype. They developed a mobile app which makes use of automated (fake) Twitter accounts to transmit secret messages. The app is named M3 - Mobile Magic Mirror. The app encodes a hidden message (a sequence of bits) through the online activities of those automated Twitter accounts rather than hiding it in the content of the message, which is how traditional information hiding systems do. The aim of this research is to assess how profile pictures and profile summaries on Twitter influence credibility judgements of the fake profiles this app uses. This can help gain insight on how to improve the information hiding technology and other related systems requiring the use of automated social media accounts.

The rest of the paper is organized as follows. The next section will give a brief overview of some technical background. Section III describes the methodology we followed for the empirical study. The results are discussed in Section IV and some further discussions are given in Section V. The last section concludes the paper with future work.

II. THEORETICAL BACKGROUND

The fake Twitter profiles created by the M3 mobile app do not have profile pictures or profile summaries. We felt profile pictures are likely to be a very important factor for the credibility of Twitter profiles. We therefore decided to manually add one facial image to each M3 Twitter account. Other two factors we considered are the profile summary and the stated gender.

This section gives some technical background and related work around the three factors that can influence credibility of social media accounts. The first two parts of this section focus on how visual (trustworthiness conveyed through a facial image as the profile picture) and verbal (trait adjectives in the profile summary) cues influence credibility judgements. The last part of this section assesses whether the gender of the person in a social media profile influences credibility judgements.

A. Face as a Cue for Credibility

As far as we know, little cyber security research has addressed how faces on social network sites impact judgements of authenticity. In particular, research has shown that people smiling in their profile pictures influence trustworthiness [8]. This personality trait has been found to play a key role in perceived source credibility [10]. Trustworthiness plays a crucial role in social judgement in general as it belongs to the morality dimension. Morality refers to correctness toward a social target such as being honest or fair which has been found to play a primary role in social perception and judgement [13,14]. For a review see [12]. In the impression formation literature, Willis and Todorov [15] suggested that people infer personality traits, including trustworthiness, from faces even after only 100-ms exposure. The correlation of these quick impressions from facial appearance with judgements made without time constraints was high (r = .73 and, when controlling for attractiveness, r = .63). People's sensitivity to subtle cues of trustworthiness from faces is enhanced when people wish to protect themselves [16]. Research from a neurobiological perspective has also suggested that facial trustworthiness judgements are part of a threat based system which responds specifically to perceived emotions of happiness and anger [17].

The present study will build on this literature by using real faces that naturally vary by facial trustworthiness as a function of the displayed emotions of happiness and anger. Indeed, happiness and anger have been associated with trustworthiness and untrustworthiness, respectively [18].

B. Influence of Verbal Cues on Credibility Judgement

As mentioned earlier, people use profile summaries as a source of information for forming impressions [8]. In the computing domain, only a few studies have looked into the effects of words on credibility judgements. A study on online daters' profiles found that those with a longer "about me" section were seen as more trustworthy [19] as it theoretically reduces uncertainty. This is also known as the uncertainty reduction hypothesis [8]. Toma and D'Angelo [20] found a similar effect for the quantity of words in an online health environment.

These types of studies focus more on the quantity of the words. However, precedent exists for different levels of personalization affecting credibility judgements on Twitter [21,22]. The studies found interesting effects on the types of words corresponding but more to a message rather than the authenticity of the profiles. We thus aimed to answer the question of how different types of words can influence people's authenticity judgements of social media profiles.

As outlined above, evidence has suggested that humans form impressions with particular weight being placed on information related to morality. This holds also when people read descriptions of others in scenario-based studies. For example when people have to form impressions of an unfamiliar ethnic group based on a description that includes trait adjectives balanced in their favorability [14].

C. Gender Effects on Credibility

There is little literature in the cyber community on how people interact with social bots or online entities that vary by gender.

In social impression processes McAleer and colleagues [23] found that attractiveness in male voices was correlated with dominance and attractive female voices were correlated with trust/likeability. Sutherland *et al.* [24] averaged faces on approachability and dominance. The 'high approachable' face was feminine and smiling. The 'low approachable' face seemed male and was neutrally/negatively emotional. The inverse is also supportive as females who are counter-stereotypically masculine are rated much more negatively in trust and likeability [25]. Recently, Jahng and Littau have found no significant effects of the profile owner's gender on the credibility of journalists using Twitter, although they argued that this finding might be specific to journalists [26]. Given the conflicting results in the literature, we were interested in testing the gender effects on credibility judgements of Twitter profiles.

D. Research Questions

This research addressed the following questions. First, we investigated what modality, visual versus textual, has the greatest impact on people's credibility judgements. In particular, how does trustworthiness/untrustworthiness conveyed by facial images influence Twitter profiles' credibility? How do morality-related vs competence-related trait adjectives in profile summaries influence credibility? Does the profile owner's gender influence credibility? Finally, we also wanted to investigate what part/s of the Twitter profiles users use to make their credibility judgements.

E. Hypotheses

We hypothesized that Twitter profiles that have a profile picture of someone who is happy/smiling (and thus associated with trustworthiness) will have higher credibility than those with angry/annoyed faces (and thus associated with untrustworthiness). Secondly, we hypothesized that Twitter profiles with a profile summary that contains morality-related

traits will have higher credibility than profiles with competence-related traits. Both hypotheses are based on the social psychological literature that shows the primary role of morality in social perception and judgement [12]. Given the conflicting results in the literature, our hypothesis about any gender effect on the credibility of Twitter profiles was explorative.

III. METHODOLOGY

A. Favorable Ethical Approval

The study received a favorable ethical opinion from the Faculty of Health and Medical Sciences Ethics Committee Ethics Committee at the University of Surrey (Ref.: FT-PSY-287-16).

B. Participants

Seventy-five participants (72 were first and second-year Psychology students at the University of Surrey who received a lab token for their participation) took part in the pre-test to select the traits to be included in the summary profiles. There were 68 females and 7 males (M_{age} = 19.39, SD = 2.9, range: 17-37).

Twenty-four participants took part in the experimental study (14 males and 10 females, M_{age}: 30.13, SD = 11.25, range = 21-58). The sample was recruited through convenience sampling. It was heterogeneous: 11 participants

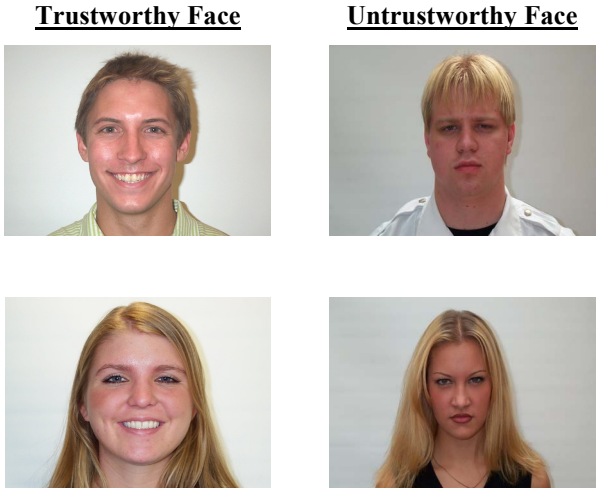

Fig. 1. Sample trustworthy and untrustworthy faces taken from the face database by Minear and Park [27].

were students, whereas the other recorded professions (e.g., "marketing", "data analyst", "teacher") involved familiarity with computers and they all spent time on social media.

C. Design

We used a 2 (face trustworthiness: trustworthy vs untrustworthy) X 2 (summary profile: morality- vs competence-related traits) X 2 (gender of Twitter profile's owners: male vs female) within-subjects design.

D. Face and Trait Selection

a) Face Selection Pre-Test

We selected the profile pictures from the face database created at the University of Michigan by Minear and Park in 2004 [27] (see Fig. 1). We conducted a pre-test with 10 participants (2 male and 8 female, M_{age} = 29.4, SD_{age} = 11.25, range: 23-58). Participants judged "to what extent do you believe the person in this picture is trustworthy?" on a 7-point Likert scale (1 = not at all, 7 = extremely) for 56 photos with different facial expressions (from angry/annoyed, neutral to happy). Given the findings by Todorov *et al.* [18], we predicted that the persons in the pictures who looked angry/annoyed would be trusted the least and those who looked happy would be trusted the most.

A Friedman test was used to analyze the levels of trust. The results showed that the effect of expression was significant, $X^2(2, N = 10) = 20$, $p < .001$. Pairwise comparisons, conducted by means of Wilcoxon tests, revealed a significant difference between the angry/annoyed (Mdn = 3.04) and neutral expressions (Mdn = 3.89, Z = -2.80, p = .005, r = .63) and the neutral expression (Mdn = 3.89) and the happy expression (Mdn = 5.30, Z = -2.80, p = .005, r = .63).

We selected 4 trustworthy and 4 untrustworthy male faces and 4 trustworthy and 4 untrustworthy female faces. We averaged participants' scores across gender and trustworthiness and compared them against the midpoint. The selected faces were symmetrically polarized in terms of trustworthiness as the final trustworthiness score (Mdn = 1.44) did not differ from the reversed untrustworthiness scores when compared relative to the midpoint (Mdn = 1.63, Z = -.652, p = .514, r = .15). They were also symmetrically balanced in terms of gender with males (Mdn = 1.00) not differing significantly from females when compared relative to the midpoint (Mdn = 1.31, Z = -.085, p = .932, r = .29).

b) Trait Selection

We also pre-tested the traits to be included in the summaries. The initial set comprised of 45 pairs (e.g., friendly-unfriendly) of traits (15 pairs per dimension, i.e., morality, sociability, competence).

It was conducted through an online questionnaire (using SurveyMonkey and advertised on SONA). Participants were asked two types of rating. First, they were asked to rate to what extent each characteristic referred to the morality/immorality, competence/incompetence and sociability/unsociability domains on scales from 1 (*little*) to 7 (*a lot*). Then they were asked, for each trait, a valence rating on a scale from -3 (*very negative*) to 3 (*very positive*) with 0 (*neutral*) as the midpoint.

We selected the following positive, morality-related traits: righteousness, loyalty, charitableness, virtuousness, fairness, altruism, respectfulness and forgivingness. The positive, competence-related traits included: intelligence, wisdom, competent, skillfulness, capability, organization, efficiency and preparedness.

We then averaged the scores which were analyzed by means of paired *t*-tests. The *t*-test for valence yielded a non-significant result, $t(74)= -1.233$, $p = .221$, $r = .14$, CI [-0.2, 0.48], showing that the valence was not different between the morality ($M = 5.77$, $SD = .57$) and competence domains ($M = 1.84$, $SD = .57$) which excluded any potential undue influence of valence. The *t*-test for relatedness (to their respective dimensions) yielded a significant result, $t(74)= 2.387$, $p = .02$, $r = .27$, CI [0.034, 0.38], between morality ($M = 5.77$, $SD = 1.05$) and competence ($M = 5.97$, $SD = 1.06$) suggesting that competence traits were judged as more prototypical compared with morality traits. However, as the answers were recorded on a 1-7 Likert scale, it was felt a difference in mean scores of 0.2 was deemed not to represent an extreme categorical difference.

As a potential confound it was assessed that morality-related traits were unrelated to the competence domain and vice versa. One-sample *t*-tests against the scale midpoint of 4 found that the moral traits ($M = 3.32$, $SD = 1.36$) were significantly unrelated to the competence domain, $t(74) = -4.214$, $p < .001$, $r = .44$, CI [-0.99, -0.36], and the competence traits ($M = 2.81$, $SD = 1.38$) were significantly unrelated to the moral domain, $t(74) = -7.455$, $p < .001$, $r = .65$, CI [-1.51, -0.87]. The results suggest that no changes needed to be made for the final lists of moral and competence traits.

E. Profile Summary Length

Profile summary length was consistent across profiles to avoid the confound that longer summaries would be rated as more trustworthy [19]. The range of summary length was between 12-22 words. Examples include "I'm too loyal for my own good sometimes but oh well, I'm Xena by the way!" and "Hey up I'm Dallas, I like to be fair to everyone that I meet, it is the most important thing for me".

F. Dependent Variables

We asked participants to provide three different types of judgements about the Twitter profiles presented to them. First, we asked participants a credibility judgement ("To what extent do you believe this profile to be a credible Twitter profile?") on a Likert scale from 1 (*Not at all credible*) to 7 (*extremely credible*). Then participants were asked to report their confidence in their judgements ("Regarding your answer to the previous question, to what extent are you confident in your choice?") on a 1 (*Not at all confident*) to 7 (*Highly confident*) scale. Finally, participants answered an open-ended question on the parts of the profiles that led most to the credibility judgements ("Regarding your answer to the first question on credibility, please indicate which part or parts of the profile led you most to your credibility judgement").

G. Procedure

Participants were welcomed to the laboratory and sat at a desktop PC. It displayed the information sheet and consent form on the online survey builder website Qualtrics. Qualtrics randomly presented blocks of questions. At the start of every block was the name of the fake Twitter profile which corresponded to a Word file containing the web links to the Twitter profiles in which the participant clicked on to be taken straight to it. Participants gave judgements on 16 fake Twitter profiles in total. Participants were given 60 seconds to view the profile and warned at the 45 second mark, as timed by the experimenter on a mobile phone stop watch. Once 60 seconds had passed the participant closed down the web page with the profile on it. For each profile, participants were then asked to record how credible they thought the profile was and how confident they were in their credibility judgements. The last question in the block was open-ended and asked the participants to identify the parts of the profile that led them to their decision. Like Wang *et al.*' study [7], a question asking participants to state their experience of social media networks was included at the end ("What is your experience with online social networks?" and given the options of 'None at all',' 0-2 years', '2-5 years', '5-10 years' and '10+ years'). Participants were also asked "How much do you use social media per day in terms of hours spent?" which was open-ended. Finally they were asked for basic demographics, were paid £5 in cash and debriefed.

H. Profile content and the M3 App

The app used to generate the original fake profiles is the "M3 – Mobile Magic Mirror App". This app (developed by the Surrey Centre for Cyber Security) is designed to transmit

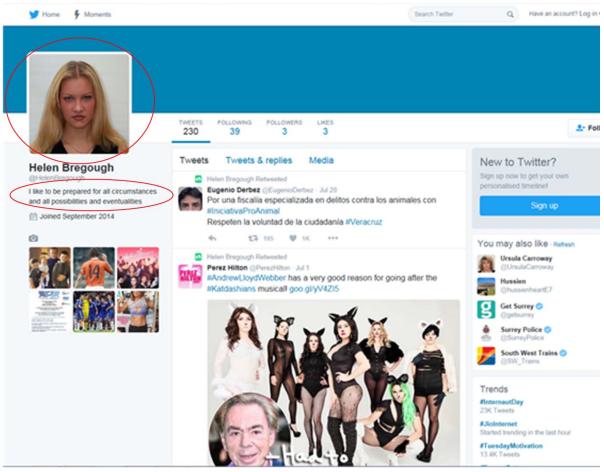

Fig. 2. Example Twitter profile. The face and profile summary that we manipulated within participants are circled in red.

hidden information through social media activity on 32 fake Twitter profiles. Those profiles were then manipulated by adding their profile pictures, profile summaries and profile owners' gender (the latter was conveyed by means of the names and facial appearance in the profile pictures).

IV. RESULTS

The data were screened for the test assumption of normality. We then analyzed the main effects and interaction effects of credibility and confidence scores. The open-ended questions were codified, assessed for randomness of missing data and percentages of the codes were calculated for general comparisons.

A. Normality Assumptions Check

The *z*-scores of the skewness and kurtosis statistic for the dependent variables all fell between +/- 0.13 to +/- 1.87 which all fall below the threshold of +/- 1.96 suggested to achieve normal distribution with small sample sizes [28].

B. Analysis of the Credibility Judgements

A grand mean credibility score was computed by averaging the credibility scores across all the conditions for all the participants. A one-sample *t*-test showed that the grand mean of credibility (M = 3.07, SD = 1.27) was significantly lower than the scale midpoint of 4, $t(23) = -3.572$, $p = .002$, $r = .60$, 95% CI [-1.46, -0.39]. This suggests that regardless of the manipulations of the independent variables, overall the profiles were perceived as lacking credibility (see Fig. 3)

A repeated-measures ANOVA on the 1-7 credibility judgements showed that there were no main effect of facial trustworthiness, $F(1, 23) = .326$, $p = .573$, $\eta^2 = .001$, profile summary, $F(1, 23) = .446$, $p = .511$, $\eta^2 = .004$, or gender, $F(1, 23) = .504$, $p = .485$, $\eta^2 = .004$.

The ANOVA also yielded no significant interaction effects of facial trustworthiness by profile summary, $F(1, 23) = .199$, $p = .659$, $\eta^2_p = .009$, of facial trustworthiness by gender, $F(1, 23) = .037$, $p = .848$, $\eta^2_p = .002$, of profile summary by gender, $F(1, 23) = 1.653$, $p = .211$, $\eta^2_p = .067$, and the three-way interaction of facial trustworthiness by profile summary by gender, $F(1, 23) = 1.866$, $p = .185$, $\eta^2_p = .075$.

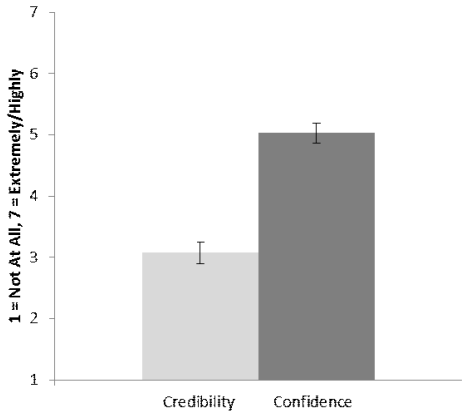

Fig. 3. Grand means of credibility and confidence scores. Error bars represent 95% confidence intervals.

C. Analysis of the Confidence Judgements

A grand mean of confidence was computed by averaging confidence scores across conditions for all participants. A one-sample *t*-test showed that the grand mean of confidence was significantly higher than the scale midpoint of 4 (M = 5.03, SD = 1.18), $t(23) = 4.27$, $p < .001$, $r = .66$, 95% CI [-0.53, 1.53]. This along with the grand mean of credibility, suggests that the fake profiles were unconvincing and that participants were highly confident in their credibility judgements (see Fig. 3).

A repeated-measures ANOVA showed there were no main effects of facial trustworthiness, $F(1, 23) = 2.484$, $p = .105$, $\eta^2 = .003$, or gender, $F(1, 23) = .423$, $p = .522$, $\eta^2 = .0004$, in the confidence participants had in their judgements. However, there was a significant main effect of profile summary showing that morality-related traits produced significantly lower confidence ratings (M = 3.03, SD = 1.22) than competence traits (M = 5.03, SD = 1.23), $F(1, 23) = 25.581$, $p < .001$, $\eta^2 = .437$. We also compared the mean confidence scores in the two conditions against the scale midpoint of 4. When the profile summaries included morality-related traits, the participants' confidence was significantly lower than the scale midpoint, $t(23) = -3.910$, $p = .001$, $r = .63$, 95% CI [-1.49, -0.46]. When the profile summaries included competence-related traits, the participants' confidence was significantly higher than the scale midpoint, $t(23) = 4.086$, $p < .001$, $r = .65$, 95% CI [0.51, 1.55]. Put together, these results suggest that competence-related traits served to increase participants' confidence in their credibility judgements whereas morality-related traits decreased it (see Fig. 4).

A repeated-measures ANOVA yielded no significant interaction effects of facial trustworthiness by profile summary, $F(1, 23) = .159$, $p = .694$, $\eta^2_p = .007$, of facial trustworthiness by gender, $F(1, 23) = 1.045$, $p = .317$, $\eta^2_p = .043$, of profile summary by gender, $F(1, 23) = 1.201$, $p = .284$, $\eta^2_p = .05$, or the three-way interaction of facial trustworthiness by profile summary by gender, $F(1, 23) = .201$, $p = .658$, $\eta^2_p = .009$.

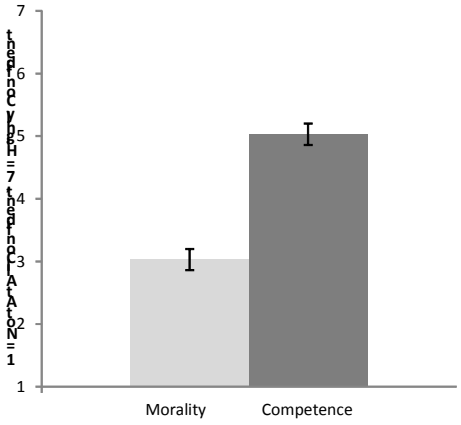

Fig. 4. Mean confidence scores for morality- and competence-related traits. Error bars represent 95% confidence intervals.

D. Analysis of Open-Ended Questions

Each open-ended response was scored "1" if it contained a part of the profile, and "0" if not. There were four code types. Code 1 was for mentioning the face/profile picture (e.g., "Why would you have a profile picture like that?" and "Profile photo looks natural"). Code 2 was for mentioning anything related to the profile summary which contained the trait manipulation (e.g., "the sentence below the profile picture" and "Only a righteous person would say being so is important to them"). Code 3 was for mentioning anything to do with the content of the Twitter profile that was unrelated to the independent variables (e.g., the joining date, extra photos, the tweets, number of followers: "Content", "no tweets", "spammy followers", "Joined October 2014 like the rest", "random

photos…"). Code 4 was for anything mentioned outside of the previous three codes (e.g., "Insti(n)ct", "I'm not sure", "They can't all be fake, the choice is getting harder", "It just doesn't feel right, not honest").

Using these four codes accounted for answers with references to multiple parts of the profile. For instance, one participant answered: "Unus(u)al profil(e) picture, only retweets, platitude for bio, arbit(r)ary suggestions" received a '1' in code 1 for mentioning the profile picture, a '1' in code 2 for the profile summary (also known as bio), a '1' in code 3 for content and a '0' for code 4.

After the coding was complete, the data were analyzed for missing values and frequencies of each response in each condition were calculated.

a) Missing Values

Out of 384 potential open-ended answers, 15 (8.2%) were missing. Little's Missing Completely at Random Test was not significant, χ^2 (97, N = 384) = 99.78, p = .403. This result shows that the data were missing randomly and not as a result of a confounding effect.

b) Frequency Analysis

The 369 recorded responses contained 470 positive codes (see Table 1). The content of the Twitter profiles (e.g., number of followers, joining date), which is what the M3 app relies on, was mentioned more than any other parts of the profile across conditions and ranged from 54.2% to 75%.

TABLE I. DESCRIPTIVES OF OPEN-ENDED QUESTIONS

Open-Ended Responses		
Code	***Counts***	***% of responses***
Profile picture (face)	71	15
Profile summary (bio)	67	14.3
Content	249	53
Other	83	17.7
Total	470	100

Frequencies of codes were compared across each of the independent variables. This was done to complement the previous analyses we conducted on credibility and confidence judgements which considered those independent variables. Furthermore, we wanted to assess the changes of codes within the context of the other codes. For instance, it is not enough to know how the use of the profile picture increases, it is important to know if the rise comes at the expense of profile summaries. A series of contingency chi-square tests were performed on all the divisions (i.e., by face, traits, and gender). There were not significant effects of trustworthy vs untrustworthy faces, χ^2 (3, N = 470) = 2.463, p = .487, morality- and competence-related traits in the profile summaries, χ^2 (3, N = 470) = 5.464, p = .110, or gender of the profile's owner, χ^2 (3, N = 470) = 1.27, p = .737.

V. DISCUSSION

This study aimed to investigate the psychological mechanisms underlying credibility judgments of the authenticity of Twitter profiles. To do so we used fake profiles created by the M3 app [11], an application which generates only retweets. The results showed that overall participants were able to notice that the Twitter profiles were inauthentic. Furthermore, they had relatively high confidence in their judgements (see Fig. 3). Confidence judgements were included to provide a more subtle measure of participants' trust. This dependent variable revealed the most interesting finding of this research. Participants trusted more their judgements when the profile summaries included competence-related traits. Competence is one of the fundamental dimensions used to make social judgements [12,29].

Contrary to our hypotheses, profile pictures, trait adjectives in profile summaries, and the profile owner's gender did not influence participants' credibility judgements. In particular, our first hypothesis was not supported as there was no evidence of increased credibility with trustworthy vs untrustworthy faces. The second hypothesis was not supported either as there was no evidence of increased credibility when the profile summary contained morality-related vs competence-related traits. Thus, although participants were able to detect that the profiles lacked credibility, their explicit judgements were not influenced by the variables we manipulated. This finding could be explained by the data from the open-ended questions. In their self-reports, participants mostly mentioned the content of Twitter profiles, which is what the M3 app relies on, rather than the other parts of the profiles which included the information we manipulated.

There are three issues when accepting the non-significant results in this study. The first issue is the lack of statistical power to detect effects due to the small sample size [28,30]. Small sample sizes attenuate the power of statistical analyses to find effects that are real. The second issue is the wording of the credibility question. Most previous research has specifically asked participants to indicate levels of "trust" in a face [17,24,15] rather than asking for credibility judgements of a whole profile. It may be that using the word 'trust' evokes a greater sense of personal risk, and self-protection motives have been shown to increase facial trustworthiness judgements [16]. The third issue concerns the overall low credibility. This result shows the fake Twitter profiles created by the M3 app appear to be unconvincing. This could be due, at least partly, to the fact that the Twitter accounts did not generate any original message. Future research could address the issue of people's credibility of Twitter profiles by using fake profiles that contain more genuine tweets instead of retweets.

The analysis of the open ended questions showed that the information we coded as "content" (code 3, see Table 1) was dominant in the credibility process. This may also be a conservative estimate because the coding system did not allow for the presence of multiple references to different parts of the

content. The coding system had other issues as well. The first is that for content, there were answers related to more automatic, implicit processes which are not accounted for by the coding system which merely notes whether a specific content is explicitly mentioned. The second issue with the coding system is that it assumed components of the profile are equal in their weight in the decision-making process in the presence and/or absence of other components. For instance, is the influence of the profile picture stronger when it is the only mentioned component in the credibility judgement or is its influence decreased in the presence of other components such as the content, along the lines of a dilution effect [31]? One further problem is that the codes do not indicate whether the component mentioned decreased or increased the credibility score. To counteract these issues further studies may wish to ask participants to rate to what extent and in what direction each of the mentioned components were in the decision making process.

A. *Future Directions*

The open-ended questions hold insight into potential improvements of the M3 app. Participants perceived the "random" nature and "inconsistency in content" in the retweet activity which is more characteristic of a (ro)bot than a person with passions and interests. One participant noted: "Unrealistic, no sense of someone real". Given the preliminary nature of this research, we have focused on fake Twitter profiles and not on genuine profiles with real tweets or incorporating real tweets into the fake profiles. Future research could systematically investigate the credibility attribution process with more ecologically valid materials. All the fake Twitter profiles drew from the same Twitter sources for their Tweets, such as celebrities, sports stars, large companies etc. A method to increase people's credibility could involve disseminating different types of Twitter sources across the profiles. In particular, the information source type could be manipulated. For instance, one of the fake profiles could draw from music artists to give the impression of being a music fan, the same can be done with sports stars, technology companies, fashion etc. Participants also realized there were no individual tweets, but only retweets. On social media, it is odd to merely be a conduit for other people's opinions which is why only having retweets could undermine the credibility of the profile, as we previously noted. This can be rectified by creating a small pool of responses to be used as actual tweets.

Individual differences in people's credibility judgements and practice effects are also avenues for further study. Finally, future studies could further deepen this strand of research by analyzing possible differences between implicit measures and explicit self-reports of sources of information for credibility judgements. In this sense, the use of eye-trackers could elucidate what features in a profile are attention-grabbing and attention-holding.

VI. Conclusion

Alan Turing asked the question of whether a human being could make as many mistakes in deciding who is a man and who is a woman in an "imitation game" played by a machine rather than by other human beings [32]. This question is still timely nowadays. Interactions with machine-created profiles are the subject of research [7]. The present study suggests that fake Twitter profiles generated by the mobile app are not entirely trusted by human being. This finding points to the overall ability of human beings to detect authenticity of online social media. Future research could nail down what factors drive human beings' implicit and explicit judgements of credibility, and how that can help the use of automated (fake) social media accounts for legitimate purposes (e.g., using the M3 app to achieve secure communications in a nation with governmental Internet censorship).

Acknowledgments

The work of the second and the third authors was partly supported by the UK part of a joint Singapore-UK research project "COMMANDO-HUMANS: COMputational Modelling and Automatic Non-intrusive Detection Of HUMan behAviour based iNSecurity", funded by the Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/N020111/1.

References

[1] D.B. Clark, "The bot bubble: how click farms have inflated social media currency," The New Republic, vol. 246, pp. 32-41, 2015.

[2] NSPCC (2015) "Always there when I need you": ChildLine review: what's affected children in April 2014 - March 2015. Retrieved September 5, 2016, from https://www.nspcc.org.uk/globalassets/documents/annual-reports/childline-annual-review-always-there-2014-2015.pdf

[3] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: threats and solutions," IEEE Communication Surveys & Tutorials, vol. 16, pp. 2019-2036, 2014

[4] A. Caspi, and P. Gorsky, "Online deception: prevalence, motivation, emotion," Cyberpsychol Behav, vol. 9, pp. 54-59, 2006.

[5] P.N. Howard, and B. Kollanyi, "Bots,# StrongerIn, and# Brexit: computational propaganda during the UK-EU referendum," Retrieved September 5, 2016 from http://ssrn.com/abstract=2798311

[6] C. Edwards, A. Edwards, P.R. Spence, and A.K. Shelton, "Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter," Comput Human Behav, vol. 33, pp. 372–376, 2014. http://doi.org/10.1016/j.chb.2013.08.013

[7] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B.Y. Zhao, "Social turing tests: crowdsourcing sybil detection," 2012. Retrieved September 5, 2016, from http://arxiv.org/abs/1205.3856

[8] C.L. Toma, " Counting on friends: cues to perceived trustworthiness in Facebook profiles," Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, pp. 495-504, May 2014.

[9] D. Westerman, P.R. Spence, and B. Van Der Heide, "A social network as information: the effect of system generated reports of connectedness on credibility on Twitter," Comput Human Behav, vol. 28, pp. 199-206, 2012.

[10] D. Westerman, P.R. Spence, B. Van Der Heide, "Social media as information source: recency of updates and credibility of information," J Comput Mediat Commun, vol. 19, pp. 171-183, 2014.

[11] S. Li, and A.T.S. Ho, "Hiding Information in a Digital Environment", Publication No. WO/2016/075459, International Application No. PCT/GB2015/053412, GB application filed on 11th November 2014, international application filed on 11th November 2015, published by WIPO (World Intellectual Property Organization) on 19th May 2016

[12] M. Brambilla, and C.W. Leach, "On the importance of being moral: the distinctive role of morality in social judgment," Social Cognition, vol. 32, pp. 397–408, 2014. http://doi.org/10.1521/soco.2014.32.4.397

[13] M. Brambilla, P. Rusconi, S. Sacchi, and P. Cherubini "Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering," Eur J Soc Psychol., vol. 41, pp. 135–143, 2011. doi:10.1002/ejsp.744

[14] M. Brambilla, S. Sacchi, P. Rusconi, P. Cherubini, V.Y. Yzerbyt, "You want to give a good impression? Be honest! Moral traits dominate group impression formation," Br J Soc Psychol., vol. 51, pp. 149–166, 2012. doi:10.1111/j.2044-8309.2010.02011.x

[15] J. Willis, and A. Todorov, "First impressions: making up your mind after a 100-Ms exposure to a face," Psychol Sci, vol. 17, pp. 592–598, 2006. http://doi.org/10.1111/j.1467-9280.2006.01750.x

[16] S.G. Young, M.L. Slepian, and D.F. Sacco, "Sensitivity to perceived facial trustworthiness is increased by activating self-protection motives," Soc Psychol Personal Sci, vol. 6, pp. 607–613, 2015. http://doi.org/10.1177/1948550615573329

[17] A.D. Engell, A. Todorov, and J.V. Haxby, "Common neural mechanisms for the evaluation of facial trustworthiness and emotional expressions as revealed by behavioral adaptation," Perception, vol. 39, pp. 931–941, 2010. http://doi.org/10.1068/p6633

[18] A. Todorov, C.P. Said, A.D. Engell, and N.N. Oosterhof, "Understanding evaluation of faceson social dimensions," Trends Cogn Sci, vol. 12, pp. 455–460, 2008.

[19] C.L. Toma, and J.T. Hancock, "What lies beneath: the linguistic traces of deception in online dating profiles", Journal of Communication, vol. 62, pp. 78-97, 2012.

[20] C.L. Toma, and J.D. D'Angelo, "Tell-tale words: linguistic cues used to infer the expertise of online medical advice," J Lang Soc Psychol, vol. 34, pp. 25–45, 2015. http://doi.org/10.1177/0261927X14554484

[21] M.R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing? Understanding microblog credibility perceptions", Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp. 441-450, 2012.

[22] G. Yilmaz, and J.M. Quintero Johnson, "Tweeting facts, Facebooking lives: the influence of language use and modality on online source credibility," Commun Res Rep, vol. 33, pp. 137–144, 2016. http://doi.org/10.1080/08824096.2016.1155047

[23] P. McAleer, A. Todorov, and P. Belin, "How do you say "Hello"? Personality impressions from brief novel voices," PLoS ONE, vol. 9, e90779, 2014. http://doi.org/10.1371/journal.pone.0090779

[24] C.A.M. Sutherland, J.A. Oldmeadow, I.M. Santos, J. Towler, D.M. Burt, and A.W. Young, "Social inferences from faces: ambient images generate a three-dimensional model," Cognition, vol. 127, pp. 105–118, 2013. http://doi.org/10.1016/j.cognition.2012.12.001

[25] C.A.M. Sutherland, A.W. Young, C.A. Mootz, and J.A. Oldmeadow, "Face gender and stereotypicality influence facial trait evaluation: counter-stereotypical female faces are negatively evaluated," Br J Psychol, vol. 106, pp. 186–208, 2015. http://doi.org/10.1111/bjop.12085.

[26] M.R. Jahng, and J. Littau, "Interacting is believing: interactivity, social cue, and perceptions of journalistic credibility on Twitter," Journal Mass Commun Q, vol. 93, pp. 38-58, 2016.

[27] M. Minear, and D.C. Park, "A lifespan database of adult facial stimuli," Behav Res Methods Instrum Comput, vol. 36, pp. 630-633, 2004.

[28] A.P. Field, Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock "n" Roll (4th edition). Los Angeles: Sage, 2013.

[29] S.T. Fiske, A.J.C. Cuddy, and P. Glick, "Universal dimensions of social cognition: warmth and competence," Trends Cogn Sci, vol. 11, pp. 77-83, 2007.

[30] S.M. Quintana, and S.E. Maxwell "A Monte Carlo comparison of seven ε-adjustment procedures in repeated measures designs with small sample sizes," J Educ Behav Stat, vol. 19, pp. 57-71, 1994. http://doi.org/10.2307/1165177

[31] R.E. Nisbett, H. Zukier, and R.E. Lemley, "The dilution effect: nondiagnostic information weakens the implications of diagnostic information," Cogn Psychol, vol. 13, pp. 248-277, 1981.

[32] A.M. Turing, "Computing machinery and intelligence," Mind, vol. 59, pp. 433-460, 1950.