How to Trick AI: Users' Strategies for Protecting Themselves from Automatic Personality Assessment

Sarah Theres Völkel[1], **Renate Häuslschmid**[2,1], **Anna Werner**[1], **Heinrich Hussmann**[1], **Andreas Butz**[1]
[1]LMU Munich, Munich, Germany, [2]Madeira Interactive Technologies Institute, Funchal, Portugal
sarah.voelkel, renate.haeuslschmid, hussmann, butz@ifi.lmu.de
anna.werner@campus.lmu.de

ABSTRACT

Psychological targeting tries to influence and manipulate users' behaviour. We investigated whether users can protect themselves from being profiled by a chatbot, which automatically assesses users' personality. Participants interacted twice with the chatbot: (1) They chatted for 45 minutes in customer service scenarios and received their *actual* profile (baseline). (2) They then were asked to repeat the interaction and to disguise their personality by strategically tricking the chatbot into calculating a *falsified* profile. In interviews, participants mentioned 41 different strategies but could only apply a subset of them in the interaction. They were able to manipulate all Big Five personality dimensions by nearly 10%. Participants regarded personality as very sensitive data. As they found tricking the AI too exhaustive for everyday use, we reflect on opportunities for privacy-protective designs in the context of personality-aware systems.

Author Keywords

chatbot; automatic personality assessment; personality

CCS Concepts

•Human-centered computing → Empirical studies in HCI;

INTRODUCTION

> *'Personality has power to uplift, power to depress, power to curse, and power to bless.'*
> Paul Harris, Advocate & Founder of the Rotary Club.

Personality is something very personal and public at the same time. We continuously express it in our language, actions and emotions [51] and yet we react very sensitively if others reject us due to our personality [25]. When artificial intelligence is utilised to measure people's personality in numbers, we open up a delicate space for classifying, discriminating and manipulating people [49] – but also for boosting self-development and interpersonal understanding. In developing such systems, we have to consider future use scenarios and their social and political impacts. These systems will be utilised for a variety of purposes, some of which will be beneficial for the user – others detrimental, or beneficial only for other parties. In the United States, AI systems which automatically assess the interviewee's personality are used in job interviewing [82]. These tools are often entirely opaque and can exclude candidates due to, for example, an increased risk of mental health problems, which is associated with high values in neuroticism [58]. Such systems have also been shown to discriminate people depending on their gender [31]. While the systems' calculations may already be accurate in many cases, some individuals might be judged wrongly and disadvantaged for no reason.

Like other personal data, personality profiles may be captured without the users' awareness, out of their control, and for unknown and dubious purposes. For example, in 2018 *Cambridge Analytica* was accused to have manipulated peoples' votes in the U.S. 2016 election campaign by means of their Facebook profiles and trait-related personalisation of online adverts [4, 68]. Users perceive their personality traits as sensitive, personal data [35] and feel uncomfortable with sharing their automatically assessed personality profiles [78]. On the other hand, they feel pressured to share their profiles to avoid social sanctions [78]. People want control over their personality profiles [30] but they feel incapable of modifying or changing them [78]. In our opinion, these circumstances raise ethical and privacy concerns and call for measures to empower and protect users while it is still possible. Such systems need to inform users about automatic personality assessments and grant them control over it without having to expect sanctions. Yet, if this control is not granted, we need to empower users to protect themselves and disguise their personality profile from such systems. In this paper, we investigate strategies which are entirely in the hands of the user and do not depend on a specific, ethically correct system design.

We examined whether people are able to trick personality assessment chatbots, i.e., manipulate the system into calculating a profile different from what a non-manipulated assessment would deliver. Our user study included two interaction phases with a customer service chatbot that assesses the users' personality in the background of an inconspicuous dialog. In the first interaction, we asked participants to interact naturally with the chatbot in order to capture their *actual* personality (baseline profile). In the second interaction, participants were framed with a data privacy and protection story and were asked to try

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376877

to disguise their personality from the chatbot (*falsified* profile). In three interviews and questionnaires surrounding the interactions, we asked participants about which aspects of the chat they believed were factored in the personality assessment, and how they could disguise their *actual* personality. We collected 41 personality disguise strategies and assessed their efficacy. We also present insights into users' attitude towards automatic personality assessment, preferences in sharing their profile, and future behaviour in case such systems become pervasive.

BACKGROUND & RELATED WORK

Personality Measurement

Personality is defined as consistent and characteristic patterns of behaviour, emotion and cognition [51]. The most prominent paradigm for describing personality is the *Five-Factor Model*, also known as *Big Five* or *OCEAN* [11,50]. It comprises five traits, which predict an individual's tendencies of characteristic behaviour [13,14,16,29,36,38,48,51–53]:

- ***O**penness* to new experiences relates to intellectual curiosity, creativity, and being perceptive to art and novel stimuli.
- ***C**onscientiousness* relates to neatness, perseverance, reliability, and responsibility.
- ***E**xtraversion* relates to sociability, activity, and assertiveness in social interactions.
- ***A**greeableness* relates to friendliness, helpfulness, and cooperativeness in dealing with others.
- ***N**euroticism* relates to emotional stability, anxiety, and the frequency of experiencing negative affect.

Automatic Assessment of Personality

Personality traits are latent constructs and, thus, cannot be measured directly. Traditionally, standardised self-report questionnaires have been used to quantify personality. With the availability of extensive online data about behaviour, researchers found new ways to automatically infer users' personality traits from digital footprints [74,79]. Since linguistic cues reflect personality [61,65], early approaches tried to predict it from publicly available texts, e.g., blog entries [27,76]. These analyses primarily employed statistics of individual word use (e.g., number of personal pronouns, word count [47,76]) as well as phrases (e.g., [56]), using the computerised Linguistic Inquiry and Word Count (LIWC) text analysis [61]. Such linguistic analyses are, for example, used to assess the users' personality within a conversation with a chatbot.

Personality also manifests itself in social media use [42,62,71]. For example, extraverted individuals show a tendency to be more active on social media than introverted ones [3, 43]. Therefore, researchers have related users' self-reported personality to their social media profiles and computed predictive models [1]. These studies used diverse features, such as *likes* [42, 81] or user activity [26, 44] and social media platforms, such as Facebook [42, 71], Twitter [46, 73] and Instagram [73]. Azucar et al.'s [1] meta-review is a good source for details on the predictive power of social media footprints.

Researchers also investigated the link between smartphone use behaviour and personality [8, 32, 74]. For example, Stachl et al. [74] developed a smartphone app that logs interaction behaviour (e.g., app usage, mobility) in the background. They found distinct behavioural patterns such as relationships between communication behaviour and extraversion as well as between music consumption and openness. Yet, these systems are still limited in their accuracy and often not reliable for all dimensions [70, 74, 77]. Further approaches use image data [7] or music preferences [24, 55]. All these systems have in common that they can run hidden in the background, predominantly on already available data records, and may thereby assess the users' personality unnoticed by the user.

Personality-aware Personalisation

Personality can predict several important life outcomes on individual, interpersonal and social levels [59]. Examples include physical health [6], information seeking behaviour [33], trust in technology [21], subjective well-being [15], relationships with peers [37] and family [2], along with romantic relationships [17], academic success [41], job performance [40], and political attitudes [39]. All of these links might suggest future application areas for personality assessment systems.

The links between personality and interests [57, 66], preferences for music and entertainment [5, 24, 66], and consumption behaviour [67, 72] have been used to personalise recommender systems [22, 23, 35, 67, 69]. Similarly, online advertisers and social media sites tailor content and presentation to the user [26, 28, 42, 49].

Matz et al. [49] investigated the influence of personality-based persuasion. They exposed 3.5 million users to advertisements which were adjusted to their personality and found that advertisements tailored to users' extraversion and openness levels led to up to 40% more clicks and 50% more purchases than mismatched or unpersonalised advertisements. They concluded that personality-based targeting has the potential for digital mass persuasion, which can be used to successfully nudge people to adopt *'better'* behaviour, e.g., to increase the efficacy of health policies [34]. However, their studies also reveal the jeopardy of powerful manipulation [49].

Acceptance of Automatic Personality Assessment

While research has looked at users' attitude towards online profiling based on explicit user data, e.g. demographics, interests [63, 64], only little is known about their attitude towards implicit user characteristics such as personality [78]. Gou et al. [30] investigated N=256 participants' attitude and sharing preferences for computationally derived personality profiles. They found that more than 60% of users were willing to share their profiles in the workplace. However, they also emphasised two concerns: (1) Users are afraid that others might misinterpret their profiles, (2) want to control and modify their data.

Warshaw et al. [78] provided participants with personality profiles generated from their social media texts. The findings revealed a paradox: Participants considered their automatically assessed profiles accurate and were uncomfortable with sharing them. On the other hand, they felt pressured to share their personality profiles to avoid social sanctions. When given the possibility to change their profile before sharing, participants felt incapable of modifying their profiles due to over-trust in the system's algorithm. Furthermore, participants expressed concerns that companies could profile them without their consent or awareness [78].

Eiband et al. [19] investigated usage problems when interacting with intelligent everyday applications and reported that when a system did not perform to the users' satisfaction, they developed coping strategies to 'trick' it. Regarding personality assessment, however, it remains unclear whether users understand how their personality can be assessed automatically and if they are able to trick such systems.

CHATBOT

The commercial personality assessment chatbot Juji[1] is provided as an online tool for job interview and customer service purposes, but can be adapted to various use cases. The chatbot employs an evidence-based personality engine, building on tweets from 15 million Twitter users [82]. Based on demographic information of these users, such as profession, and stereotypically associated personality traits, it was trained to predict personality from the linguistic content of their tweets, resulting in reliable predictions for an input of approximately 1,000 words. These linguistic cues include, for example, word count, text length, and keywords.

We decided to use the Juji chatbot because it (1) is already in use and has reached a certain maturity, (2) has a good prediction rate in comparison to other approaches [82], (3) is available (for free) for academic research purposes and (4) provided us with the users' input and the system's output – as required to assess the user's strategies and their efficacy. The chatbot consists of the actual chat website and a backend for customisation including a selection of chatbot personalities, some of which have been evaluated in previous studies. We chose the chatbot personality Kai (male version of Kaya) for its well-balanced, open and friendly, but also professional conversation style [82]. In the section *Conversation Design* we describe how we designed our chatbot conversation.

RESEARCH APPROACH

Study Design

Our main question was whether people are able to trick personality assessment chatbots, i.e., disguise their *actual* personality from them. We approached this question by a within-subjects, repeated-measures lab study containing two equal interaction phases with a personality assessment chatbot. Participants were asked to (1) interact with this chatbot and thereby indirectly create their *actual* profile as a baseline and (2) try to disguise their personality by strategically tricking the chatbot into calculating a *falsified* profile. We also investigated participants' mental model of the chatbot and strategies to deceive it in three interviews and three questionnaires. Figure 1 illustrates the overall procedure of our lab study.

Customer Service Use Case

For the two interactions, we needed a realistic, goal-directed interaction use case that we could frame in two distinctive ways. We chose a customer service scenario mainly because of three reasons: (1) The main use case of the chatbot we utilised are customer service and job interviews. (2) Chatbots are already widely in use for customer services and, hence, likely to have been encountered by our participants in this scenario before. (3) People approach customer services with specific goals (getting help with something). Choosing common products allowed us to create interaction scenarios most people can relate to. In addition, the goal of the interaction (getting help with the products) remains the same in both chatbot interactions. In contrast, an interview for a job position would never fit every participant. One chatbot interaction contained three customer service scenarios: Solving smartphone issues, booking holidays and buying a backpack. We prepared short key fact sheets to be filled by the participants prior to the first interaction in order to ensure the goal of the interaction (solving the problem or receiving a specific product) remained unchanged. Changing the goal would open up more strategies to trick the system but would render the interaction meaningless in reality.

Collection & Visualisation of Personality Profiles

For the first interaction, we instructed participants to interact naturally and honestly with the chatbot to collect an unbiased baseline profile. The second interaction was framed by a story in which participants had to imagine interacting with a chatbot that tries to assess their personality for unknown purposes and out of their control. We explicitly asked them to try to prevent the chatbot from capturing their *actual* personality and induce a *falsified* profile. To increase their motivation, we offered an incentive of €50 for the participant achieving the largest mismatch between the two profiles.

The chatbot's web interface delivers a spreadsheet with values ranging between 0% and 100% for the Big Five personality traits and six sub-facets each. Since these tables are not interpretable by participants, we translated them into PDF documents which are understandable to non-experts. We let the chatbot calculate participants' profiles after each profile and used a Python script to visualise the resulting Big Five traits including six sub-facets each in bar charts. Textual descriptions informed participants about their meaning. Since these profiles constitute private information, the experimenter could only see them when this was offered by the participant. We decided to additionally measure participants' personality by means of the German version of the Big Five personality inventory (BFI-2) questionnaire [12], which is well established in psychology. This profile was provided to participants after the study and served as a metric for the accuracy of the chatbot's profiles.

Interviews & Questionnaires

As personality assessment systems are still novel, we assumed that most people had not gained any experience with them yet, and, hence, not formed a considerate mental model of and opinion about these systems. We therefore decided to ask participants repetitively about the same topics as they gained more experience and scheduled three interviews and questionnaires around the two interactions. We chose interviews to gather deep, qualitative insights into (1) the participants' assumptions of what factors the chatbot may use to calculate their profile (mental model) and which strategies could be effective to trick the system, (2) their attitudes towards such

[1] https://juji.io

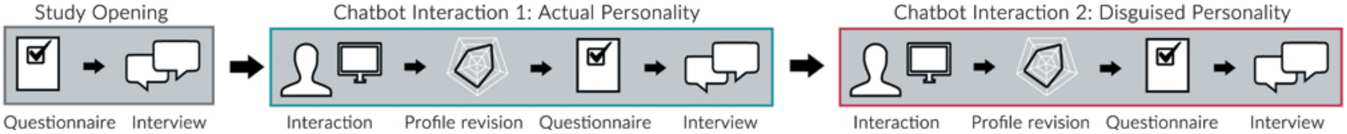


Figure 1. Our study consisted of three phases: (A) We opened the study with questionnaires and a first interview. (B) In the first chatbot interaction, our participants chatted naturally with the chatbot and reviewed their personality profile. We collected insights by means of questionnaires and an interview. (C) The second interaction was equal to the first one, but participants had to strategically manipulate the chatbot to disguise their personality.

systems and (3) their preferences to share their profile. To support these insights with directly comparable Likert scale values, we directed some questions again in questionnaires. In particular, we asked for participants' perceived accuracy of the calculated profiles and their preferences of sharing different profiles with different parties (as inspired by Gou et al. [30]). Additionally, we used a demographic and the BFI-2 questionnaire to collect information about the participants.

Conversation Design

After consultation with an employee of the chatbot manufacturer, we created a set of questions and answers for our use cases. We designed the interaction to take approximately 45-60 min, since a minimum of 1,000 words was recommended for a solid personality assessment by the Juji chatbot [82]. Several trial runs and a pilot study showed that 1,000 words roughly correspond to 45 min and that the second interaction process tended to be shorter than the first. We split the 45 min into three separate scenarios (smartphone issues, holidays booking and backpack shopping) because we considered customer service interactions of more than 15 min unrealistic. The conversation procedure and questions were designed to lead the user smoothly through an imaginary problem solving or shopping scenario. While some of the questions were focused on the product, others were more personal but still in context (e.g., *what would be ideal holidays for you?*). We assumed that these more engaging questions would motivate participants to reply with longer texts, which in turn would lead to a more accurate profile.

Study Procedure

As outlined in Figure 1, the study procedure consisted of three major phases with few sub-phases. All three phases were conducted within one study session, which lasted 2 to 3.5 h in total including short breaks between the phases. In order to ensure consistency and prevent any biasing of the participants, we prepared detailed study scripts for all instructions.

Study Opening

The experimenter asked participants to carefully read through the consent form and sign it if they agreed. She then introduced the participants to the study topic, goals and procedure. She explained that they would receive two personality profiles, which may more or less reflect their actual personality and emphasised that they were considered private information and that showing them to the experimenter was voluntary. The participants then filled out the first questionnaires and were interviewed about their initial mental model and attitude.

Chatbot Interaction 1: Actual Personality

The experimenter introduced the participants to the chatbot and the first customer service scenario. To help the participants in creating a personal case within this scenario and remaining consistent in both interactions, they had to write down the key facts in a short questionnaire (only for personal use). We requested them to adhere to their case and answer honestly to all personal questions. This procedure was repeated for the next two customer service scenarios. At the end of the first chatbot interaction, the participants received and read through their *actual* personality profile computed by the chatbot. They were requested to fill out a short questionnaire and were interviewed about their new mental model and potential strategies to disguise their personality from the chatbot.

Chatbot Interaction 2: Disguised Personality

The experimenter started the second interaction with a data privacy and protection story: The chatbot collects personality profiles for unknown purposes and participants' goal is to have the next (*falsified*) profile differ as much as possible from their *actual* profile to protect themselves.

The experimenter informed them about the financial incentive and provided the key fact sheets from the first chatbot interaction for reference. Then, the chatbot interaction, revision of the *falsified* profile and questionnaires proceeded as described in *Interaction 1*. The study was closed with another interview about participants' mental model, strategies and attitude towards personality assessment systems.

Interview Analysis

We transcribed all interviews and conducted a data-driven inductive thematic analysis with two coders. We randomly selected and independently coded six participants' interviews. We discussed each code until agreement was found and thereby merged our two coding tables into one codebook. With the codebook at hand, this procedure was repeated for another six participants. We calculated the inter-coder agreement using Cohen's κ [9]. Since participants' statements could be assigned to multiple categories, we calculated κ for each of the 195 categories using 2x2 contingency tables (code present: yes/no), similarly to [18]. We reached perfect agreement for 76.9% of the codes[2]. The remaining nine participants were split between the coders and analysed independently. We discussed all uncertainties and optimised the final codebook together.

Participants

Participants were recruited via university mailing lists. 21 participants completed the experiment (11 females; mean age 23.6 years, range 20-28 years). Participants' educational level was high (52% a-level degree, 43% university degree, 5% professional training). Participants were compensated with a

[2] For 1.0% of codes, κ was 0.67 (substantial agreement), for 6.7% of codes, κ was 0.57 (moderate agreement), and for 15% between 0.00 and 0.25 (slight agreement).

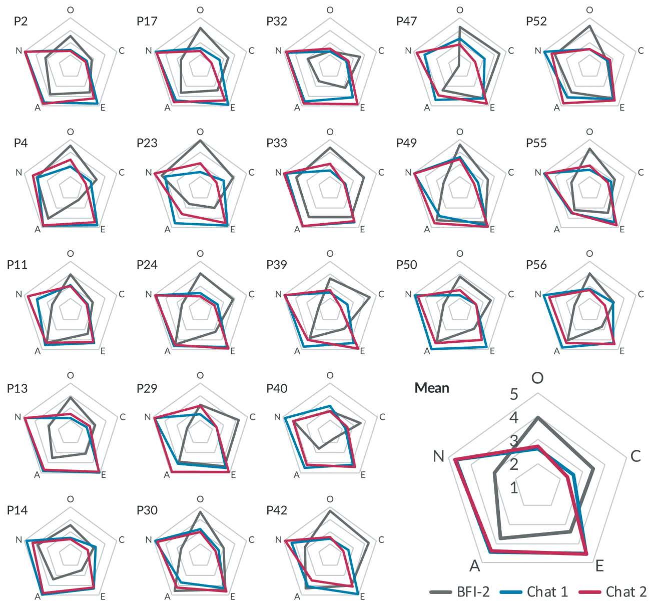


Figure 2. Each spider web shows the *BFI-2* profile (grey) as well as the chatbot-generated *actual* (blue) and *falsified* profiles (red) of one participant; the participant IDs associate the profiles to the quotes provided in the text. The large spider web shows the mean values over all participants. The spider web corners represent the Big Five personality traits.

study course credit, € 30 in cash, or a voucher. They could also keep their personality profiles, if they wanted.

Each study session was designed to last approximately 2.5 to 3 h, which we tested in a pilot study. However, we noticed large variations among participants so that each session took between 2 and 3.5 h. We scheduled 3 h time slots with participants (with 30min buffer between two participants). In case a participant did not finish within this time frame and could not stay longer than 3 h (one participant) or if it was foreseeable that they will not finish within 3.5 h (two participants), we aborted the study and excluded those participants from the analysis. In the first interaction, participants chatted between 31 and 92 minutes with the chatbot (M=57.38, SD=16.71) and used between 235 and 1,244 words (M=717.24, SD=244.72). In the second interaction, participants used between 146 and 1,725 words (M=576.86, SD=400.02). Since some participants did not close the chat window before the final interviews, we cannot report the duration of the second chat interactions.

RESULTS

Personality Profiles

Figure 2 shows participants' scores on all Big Five traits for the *actual* and the *falsified* profiles generated by the chatbot as well as the BFI-2 profile. For the comparison of the BFI-2 with the chatbot profiles, we adapted the scale of the chatbot profiles (percentages) to the BFI-2 (five-item scale). The chatbot consistently calculated high values for extraversion, agreeableness, and neuroticism but low values for openness and conscientiousness for all participants. The BFI-2 profiles show a larger variation for all traits. This resulted in discrepancies between the BFI-2 profile and the chatbot-generated *actual* profile, which were on average M=1.37 (SD = 0.60) for openness, M=0.91 (SD=0.60) for conscientiousness, M=1.17 (SD=0.77) for extraversion, M=0.87 (SD=0.59) for agreeableness, and M=1.77 (SD=0.88) for neuroticism. We calculated the Spearman's rank correlation coefficient for the relationship between the BFI-2 and *actual* chatbot-generated profile for all five dimensions but did not find any significant correlations.

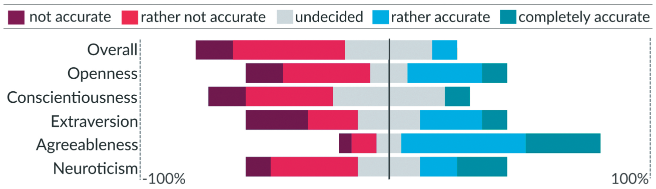


Figure 3. Participants' perceived accuracy of their chatbot generated, *actual* profile for all Big Five personality traits.

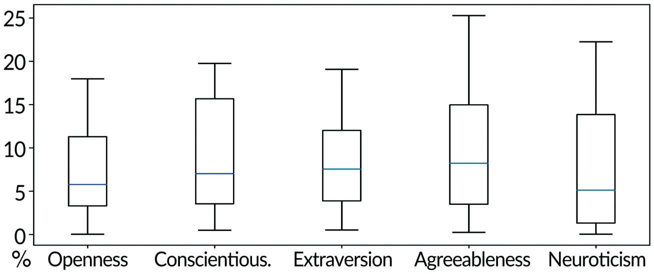


Figure 4. The boxplots show the absolute difference between the *actual* and the *falsified* chatbot generated profiles in percent.

Figure 3 shows the participants' perceived accuracy of the *actual* profile. The majority of the participants were undecided or perceived it as rather inaccurate, particularly for conscientiousness. In contrast, only agreeableness was perceived as rather or completely accurate by most participants.

On average, participants reached a difference between the two chatbot profiles of 7% to 10% for each trait (cf. Table 1). Figure 4 outlines the absolute difference for each dimension for the *actual* and *falsified* chatbot-generated personality profiles. Participants were asked to manipulate the *falsified* profile without a prescribed direction (increase or decrease). Divergences in the two directions are, hence, equally valid but might even out if not calculated as *absolute* mean difference. Our results did not match a meaningful statistical test due to the lack of a control group (same procedure but without the manipulation of the chatbot).

Personality Cues & Tricking Strategies

We asked participants what they believed a chatbot uses to assess the users' personality (*cues*). We also asked them how they could trick such a system into capturing a personality different from their *actual* personality (*strategies*). In sum, our participants suggested 214 cues and strategies. Participants made on average ten suggestions (min=5, max=20) and applied five (min=1, max=13) of them to trick the chatbot.

We aggregated highly related cues and strategies, which are based on manipulating this cue. We obtained a set of 37 distinguished cue categories and four additional, cue-independent high-level strategies. In the remainder of the paper, we refer to the cues and strategies together as *factors*.

In Figure 5, we list and explain all 41 factors and underpin them with exemplary statements from our participants. The table also shows how many participants mentioned the factors and how many applied them in the second interaction phase. Except for the high-level strategies, applying a factor means that a participant tries to vary its use, e.g. by using other keywords, showing more affection or writing shorter texts.

Participants assumed that the chatbot would factor in *specific keywords*, which the user mingles in the text (n=15), the *text length* (n=16), *topics, interests and preferences* about which the user talks (n=10) as well as the *elaborateness and detailedness* of these texts (n=11), what users report as being their (usual) *behaviour* (n=11), the amount of *provided information* and the conversation dynamics doing so (n=10), and the users' time needed to write a *response* (n=12). In contrast, the most applied strategies are varying the *text length* (n=12), *punctuation style*, *elaborateness and detailedness* of the text, and writing *opposites* to what would be replied usually (n=6 each).

We consulted the development team of the Juji chatbot in order to highlight the factors that could be effective. According to the chatbot's current implementation, all variations in the language may be effective to induce a *falsified* profile, as are all high-level strategies, given that they may impact these language features. An analysis of interaction and conversation behaviour is currently not incorporated in the chatbot's implementation and, hence, these strategies would not be effective.

Tricking in Everyday Life

Towards the end of the last interview, we asked participants whether they could imagine to apply such strategies or change their online behaviour in everyday life if personality assessment systems are employed. Our participants said they would apply such strategies in job interview situations and at work (n=11), in customer service situations (n=3), and on social media (n=2). Five participants did not see the need to trick such a system.

Participants mentioned they would change their online behaviour in order to provide less data (n=6), particularly when the user of the data is unknown. They would also stop or reduce using a system (n=5), move to alternative systems (n=3), use the systems offline (assessment is assumed to happen online) and avoid unnecessary contact (n=1 each). Participants would trick the system if the purpose of the assessment is unknown or expected to be negative (n=6) and to protect their data and profile (n=5). Some participants came up with the idea of *improving* the profile (n=3) but also concerns about anonymity and being discovered as a liar (n=3) were mentioned. For example, P47 said s/he would *'trick the system to leave a better impression in a job interview [with a chatbot]'*, however, s/he was worried about then meeting the interviewer in person and the tricking to be discovered.

As reasons why they would not try to deceive the system, participants put forward that it was too exhausting (n=10), they would forget about it (n=4), and that they *'saw no reason [as they] had no problem with [themselves]'* (n=2). Four participants mentioned that it would prevent them from receiving the benefits of using the system: For example, P23 mentioned that *'particularly with Alexa it would not be practical for me. Because I need something from her. So rather not.'* Another two participants said they do not want to be perceived in a distorted way. P13 explained, for example, that *'on social media no because I rarely post something anyway and if so I do not want it to be vitiated.'* Some participants had already given up on their data and privacy because it was already collected (n=3), they had nothing to hide (n=2) or because it would become obsolete after some time anyway (n=1). P56 mentioned, for example, that *'most online companies have so much data about me already that it does not make a difference anymore'* and P49 stated that *'the chance [of protecting the data] is long gone and one simply has to accept that. I don't think that's good but it's a lot more exhausting to disguise oneself. I don't like to do that and I have nothing to hide. I don't care.'*

Trait	***Actual* Profile**		***Falsified* Profile**		**Absolute Diff.**	
	Mean	SD	Mean	SD	Mean	SD
O	40.51	9.86	43.00	8.80	7.57	5.27
C	40.23	6.86	33.52	5.19	8.53	6.54
E	88.13	6.81	88.87	7.03	8.25	5.59
A	86.41	10.39	84.44	11.06	9.53	7.31
N	91.65	9.47	92.67	6.63	7.98	7.43

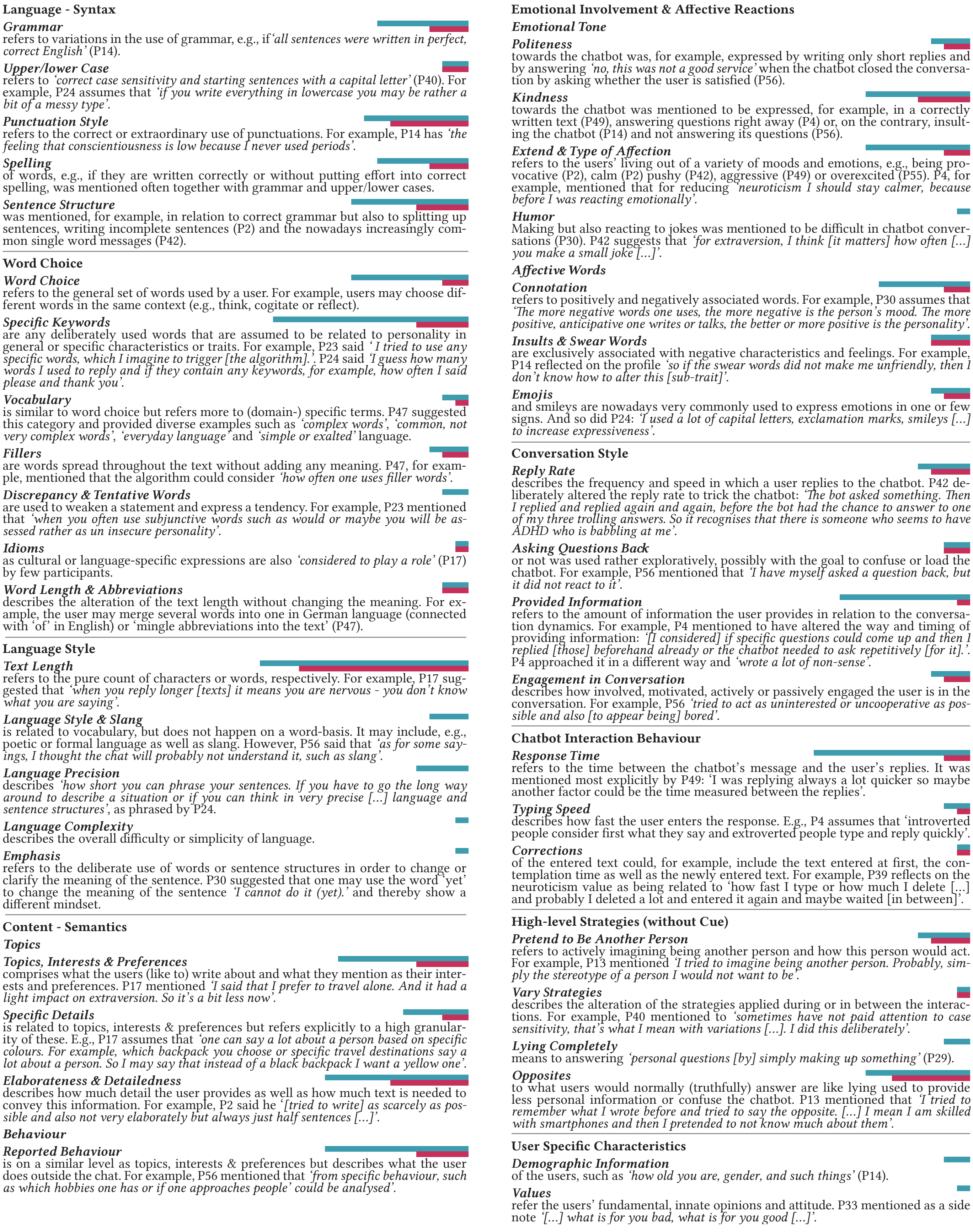
Table 1. The table shows the mean values for all traits in percentages for the *actual* and the *falsified* chatbot generated profiles. The absolute mean difference between the two profiles is depicted, which indicates how well the participants could manipulate the chatbot's assessment.

Sharing of Profiles

Figure 6 illustrates the participants' willingness to share the chatbot-generated personality profile and a *freely adapted* version, for which participants could make any changes they would like to the profile. It shows that our participants would share their *actual* and a *freely adapted* profile with family and friends but rather not with acquaintances, the public and at work. In contrast to the *actual* profile, participants' willingness to share a *freely adapted* profile slightly decreased for friends and family but slightly increased for acquaintances.

Participants' answers when and why they would share a personality profile seemed to depend on its accuracy. They would share the chatbot's *actual* profile and one that is *'corrected'* to their self-image, most prominently when they recall the experiment and discuss it with others (n=7 each). For example, P47 would ask her friends: *'That's how I was evaluated, how would you assess my character?'* However, they also mentioned they would stress the limitations of the assessment in such a conversation (n=6). Participants also said they would share their profile if it helped self-reflection (P30 & P42) and *'to understand each other better'* in team settings (P42). They also explained they would do so as they *'had nothing to hide'* (P47) or if it was advantageous in some way (P55). If it was beneficial for them, two participants would also share a *freely adapted (made-up)* profile. Nine participants found it not meaningful and eight participants underlined that it would be *'dishonest and would probably be discovered'* (P13, n=2).

Participants would not share the chatbot's *actual* or a *corrected* profile because they find it not meaningful (n=7), not accurate enough (n=1), or too private (n=6). Yet, they would share it if requested (n=3). They did not want to share it with people who



Language - Syntax

Grammar
refers to variations in the use of grammar, e.g., if *'all sentences were written in perfect, correct English'* (P14).

Upper/lower Case
refers to *'correct case sensitivity and starting sentences with a capital letter'* (P40). For example, P24 assumes that *'if you write everything in lowercase you may be rather a bit of a messy type'*.

Punctuation Style
refers to the correct or extraordinary use of punctuations. For example, P14 has *'the feeling that conscientiousness is low because I never used periods'*.

Spelling
of words, e.g., if they are written correctly or without putting effort into correct spelling, was mentioned often together with grammar and upper/lower cases.

Sentence Structure
was mentioned, for example, in relation to correct grammar but also to splitting up sentences, writing incomplete sentences (P2) and the nowadays increasingly common single word messages (P42).

Word Choice

Word Choice
refers to the general set of words used by a user. For example, users may choose different words in the same context (e.g., think, cogitate or reflect).

Specific Keywords
are any deliberately used words that are assumed to be related to personality in general or specific characteristics or traits. For example, P23 said *' I tried to use any specific words, which I imagine to trigger [the algorithm].'*. P24 said *'I guess how many words I used to reply and if they contain any keywords, for example, how often I said please and thank you'*.

Vocabulary
is similar to word choice but refers more to (domain-) specific terms. P47 suggested this category and provided diverse examples such as *'complex words'*, *'common, not very complex words'*, *'everyday language'* and *'simple or exalted'* language.

Fillers
are words spread throughout the text without adding any meaning. P47, for example, mentioned that the algorithm could consider *'how often one uses filler words'*.

Discrepancy & Tentative Words
are used to weaken a statement and express a tendency. For example, P23 mentioned that *'when you often use subjunctive words such as would or maybe you will be assessed rather as un insecure personality'*.

Idioms
as cultural or language-specific expressions are also *'considered to play a role'* (P17) by few participants.

Word Length & Abbreviations
describes the alteration of the text length without changing the meaning. For example, the user may merge several words into one in German language (connected with 'of' in English) or 'mingle abbreviations into the text' (P47).

Language Style

Text Length
refers to the pure count of characters or words, respectively. For example, P17 suggested that *'when you reply longer [texts] it means you are nervous - you don't know what you are saying'*.

Language Style & Slang
is related to vocabulary, but does not happen on a word-basis. It may include, e.g., poetic or formal language as well as slang. However, P56 said that *'as for some sayings, I thought the chat will probably not understand it, such as slang'*.

Language Precision
describes *'how short you can phrase your sentences. If you have to go the long way around to describe a situation or if you can think in very precise [...] language and sentence structures'*, as phrased by P24.

Language Complexity
describes the overall difficulty or simplicity of language.

Emphasis
refers to the deliberate use of words or sentence structures in order to change or clarify the meaning of the sentence. P30 suggested that one may use the word 'yet' to change the meaning of the sentence *'I cannot do it (yet).'* and thereby show a different mindset.

Content - Semantics

Topics

Topics, Interests & Preferences
comprises what the users (like to) write about and what they mention as their interests and preferences. P17 mentioned *'I said that I prefer to travel alone. And it had a light impact on extraversion. So it's a bit less now'*.

Specific Details
is related to topics, interests & preferences but refers explicitly to a high granularity of these. E.g., P17 assumes that *'one can say a lot about a person based on specific colours. For example, which backpack you choose or specific travel destinations say a lot about a person. So I may say that instead of a black backpack I want a yellow one'*.

Elaborateness & Detailedness
describes how much detail the user provides as well as how much text is needed to convey this information. For example, P2 said he *'[tried to write] as scarcely as possible and also not very elaborately but always just half sentences [...]'*.

Behaviour

Reported Behaviour
is on a similar level as topics, interests & preferences but describes what the user does outside the chat. For example, P56 mentioned that *'from specific behaviour, such as which hobbies one has or if one approaches people' could be analysed'*.

Emotional Involvement & Affective Reactions

Emotional Tone

Politeness
towards the chatbot was, for example, expressed by writing only short replies and by answering *'no, this was not a good service'* when the chatbot closed the conversation by asking whether the user is satisfied (P56).

Kindness
towards the chatbot was mentioned to be expressed, for example, in a correctly written text (P49), answering questions right away (P4) or, on the contrary, insulting the chatbot (P14) and not answering its questions (P56).

Extend & Type of Affection
refers to the users' living out of a variety of moods and emotions, e.g., being provocative (P2), calm (P2) pushy (P42), aggressive (P49) or overexcited (P55). P4, for example, mentioned that for reducing *'neuroticism I should stay calmer, because before I was reacting emotionally'*.

Humor
Making but also reacting to jokes was mentioned to be difficult in chatbot conversations (P30). P42 suggests that *'for extraversion, I think [it matters] how often [...] you make a small joke [...]'*.

Affective Words

Connotation
refers to positively and negatively associated words. For example, P30 assumes that *'The more negative words one uses, the more negative is the person's mood. The more positive, anticipative one writes or talks, the better or more positive is the personality'*.

Insults & Swear Words
are exclusively associated with negative characteristics and feelings. For example, P14 reflected on the profile *'so if the swear words did not make me unfriendly, then I don't know how to alter this [sub-trait]'*.

Emojis
and smileys are nowadays very commonly used to express emotions in one or few signs. And so did P24: *'I used a lot of capital letters, exclamation marks, smileys [...] to increase expressiveness'*.

Conversation Style

Reply Rate
describes the frequency and speed in which a user replies to the chatbot. P42 deliberately altered the reply rate to trick the chatbot: *'The bot asked something. Then I replied and replied again and again, before the bot had the chance to answer to one of my three trolling answers. So it recognises that there is someone who seems to have ADHD who is babbling at me'*.

Asking Questions Back
or not was used rather exploratively, possibly with the goal to confuse or load the chatbot. For example, P56 mentioned that *'I have myself asked a question back, but it did not react to it'*.

Provided Information
refers to the amount of information the user provides in relation to the conversation dynamics. For example, P4 mentioned to have altered the way and timing of providing information: *'[I considered] if specific questions could come up and then I replied [those] beforehand already or the chatbot needed to ask repetitively [for it].'*. P4 approached it in a different way and *'wrote a lot of non-sense'*.

Engagement in Conversation
describes how involved, motivated, actively or passively engaged the user is in the conversation. For example, P56 *'tried to act as uninterested or uncooperative as possible and also [to appear being] bored'*.

Chatbot Interaction Behaviour

Response Time
refers to the time between the chatbot's message and the user's replies. It was mentioned most explicitly by P49: 'I was replying always a lot quicker so maybe another factor could be the time measured between the replies'.

Typing Speed
describes how fast the user enters the response. E.g., P4 assumes that 'introverted people consider first what they say and extroverted people type and reply quickly'.

Corrections
of the entered text could, for example, include the text entered at first, the contemplation time as well as the newly entered text. For example, P39 reflects on the neuroticism value as being related to 'how fast I type or how much I delete [...] and probably I deleted a lot and entered it again and maybe waited [in between]'.

High-level Strategies (without Cue)

Pretend to Be Another Person
refers to actively imagining being another person and how this person would act. For example, P13 mentioned *'I tried to imagine being another person. Probably, simply the stereotype of a person I would not want to be'*.

Vary Strategies
describes the alteration of the strategies applied during or in between the interactions. For example, P40 mentioned to *'sometimes have not paid attention to case sensitivity, that's what I mean with variations [...]. I did this deliberately'*.

Lying Completely
means to answering *'personal questions [by] simply making up something'* (P29).

Opposites
to what users would normally (truthfully) answer are like lying used to provide less personal information or confuse the chatbot. P13 mentioned that *'I tried to remember what I wrote before and tried to say the opposite. [...] I mean I am skilled with smartphones and then I pretended to not know much about them'*.

User Specific Characteristics

Demographic Information
of the users, such as *'how old you are, gender, and such things'* (P14).

Values
refer the users' fundamental, innate opinions and attitude. P33 mentioned as a side note *'[...] what is for you bad, what is for you good [...]'*.

Figure 5. The figure shows 37 cue-based and four high-level strategies suggested by our participants in three interviews. The blue bars represent the count of participants who identified the strategy (min=1, max=16), and the red bar depicts how many participants applied it (min=0, max=13).

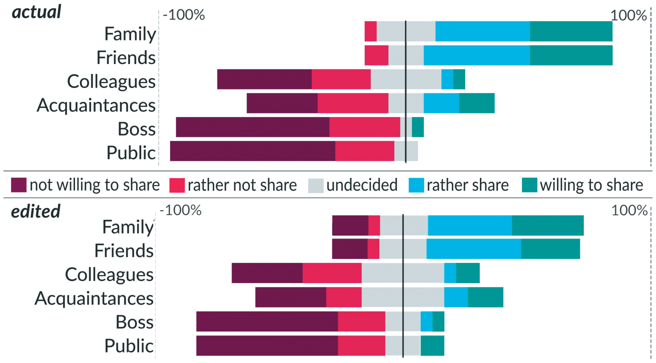


Figure 6. Participants' preferences to share their *actual* (top) and a *freely edited* (bottom) personality profiles with different parties.

do not know them (well) as they were worried about being judged (n=2), creating a wrong impression (n=5), and expected no useful feedback (n=1). For example, P30 mentioned that for a profile indicating high neuroticism, a company *'would never hire [her]. And that's awful. [...] No one has a chance then to develop if [...] you're a 100% in depression [...] that would lead to a class system'*. Three participants claimed to be reluctant to adapt the profile as they suspected a self-bias: *'I wouldn't show it the way I wish it would be because people don't assess themselves correctly'* (P33).

Attitudes & Expected Use Cases

Before and after the two interaction phases, we interviewed participants about their attitudes towards automatic personality assessment and where they expect to encounter such systems. Of our 21 participants, the majority (n=13) were initially neutral towards such developments, three had a positive attitude and five found it rather negative. After the study, two participants regarded this development more negatively; the remaining 19 participants did not change their attitude. They mentioned that their attitude towards such systems depended very much on its purpose (n=13). Slightly fewer participants expected to gain benefits from it (n=8) than being exposed to risks and disadvantages (n=10). Five participants claimed to (still) have too little knowledge to judge this development.

Most commonly, participants identified *commercial and personalisation purposes* such as personalised advertisements and recommendations in marketing (n=20) or product evaluation (n=1). In addition, they expected to encounter them in social networks (n=3), online search websites (n=2) and customer services (n=1) with the aim to increase usage time (n=1).

Personality assessment systems could *support people* when applied for medical and psychological purposes such as for psychotherapy (n=8), self-reflection and personal growth (n=3), for children's development (n=1), or for reducing exposure to stressors (n=1). For example, P42 mentioned that *'such a system, which can better reflect on one because [...] confessing the own weaknesses to one-self is not something in which the human excels, [...] could really give you another perspective.*'

They may also be used to *rate, understand, and match people*, e.g., in job interviews (n=5) and to match a team (n=2) or lovers (n=1). They expect their use in politics (n=4), research, criminology, and when effecting an insurance (n=1 each).

Participants expressed worries about protecting their *privacy* (n=5) and data (n=7) and expected that such a technology may be used to monitor people (n=3). For example, P2 said *'it goes in the direction of total surveillance [...]. It is a pretty strong instrument for control.'* In a similar vein, they expressed concerns that people might be classified (n=1), that their decisions might be anticipated or manipulated (n=2) and with disadvantages due to filter bubbles (n=3).

Participants were hardly aware that such assessments may happen without their awareness (n=4) and against their will (n=1). For example, P2 mentioned that *'it was practically like a shopping support conversation. I didn't think that this was possible. That you can assess the personality through that. I thought it would be really directed, that you know – okay, now it's about your personality. [...] And that is like a disguised personality assessment.'* They expressed a general feeling of discomfort and fear of these systems (n=6), also because these systems did not comply with human ethics (n=1). For example, P30 did not want others to have access to his/her profile because s/he *'knows, that people, malicious people or greedy people, use it for bad purposes. Against our will and against our advantage.'*

Participants stressed the low *accuracy* and limited reliability of the profiles (n=11), also because they were a stereotypical view on people (n=2) and could be manipulated (n=3). They mentioned that such systems were not sufficiently adaptable to the situation or the user (n=3) and too limited in the collected data (n=1). Yet, people acknowledged that they may reduce human errors (n=2), required resources (n=2) and pre-judgment (n=1) as well as enhance availability (n=2), efficiency (n=2) and long-term tracking (n=1). Two participants expressed their curiosity as they found them an interesting development.

DISCUSSION & LIMITATIONS

Accuracy of Automatic Personality Assessment

Previous research shows contradicting results for the profile accuracy: While Warshaw et al. [78] found automatically generated profiles from social media to be 'creepily accurate' and Youyou et al. [81] suggest that it might even surpass human judgement of personality traits, other studies show that the accuracy is often not reliable for all dimensions [70, 74, 77]. Our results also show a discrepancy between the BFI-2 profile and the first chatbot-generated (*actual*) profile. The discrepancy points at a limited accuracy, which might be caused by the word count of many participants. To create the *falsified* profile, participants often reduced the text length as a strategy. In addition, the overall study duration presented a limiting factor: If we had increased the time of a single interaction beyond one hour to increase the word count, we would have exceeded a total of 3 h by far and introduced fatigue effects.

However, since we did not want to learn about participants' personality per se but rather about the relative change in the generated profile, the absolute accuracy of these profiles was only of secondary interest. To make a valid statement about the efficacy of participants' strategies, the chatbot's implementation provides a meaningful measure. Furthermore, comparing the two generated profiles gives a first indication of their efficacy as well as of participants' capability to apply them.

We regard it as unlikely that chatbot-based job interviews significantly exceed a duration of one hour. Regarding customer service chats, we doubt that one hour will be reached at all. Consequently, when applied in these real world situations, the data collected in one interaction will most probably not be sufficient to derive a trustworthy profile with today's methods. In contrast, repeated or longer interactions with the chatbot, e.g. a frequently returning customer or a social media user, may likely improve the trustworthiness of the profile.

Efficacy of Users' Tricking Strategies

Overall, participants provided a wide variety potentially effective factors to influence the chatbot's personality assessment. The suggested strategies indicate that our participants understand how to fake agreeableness (e.g., using swear words, being (im)polite, aggressive) and extraversion better than (e.g., changing interests) other dimensions such as conscientiousness and neuroticism. They ascribed specific linguistic cues to certain personality traits, such as the use of incorrect spelling to lower conscientiousness, tentative words to insecurity, and politeness to agreeableness. This confirms previous research that people's implicit folk theories are often accurate [54]. Linguistic links less represented in folk theory, such as articles or prepositions [80], were not named by participants. However, participants suggested interaction behaviour and general conversation style (e.g., response time) as effective strategies, which are currently not included in the chatbot's analysis.

Despite the number of identified factors, participants were not very successful in tricking the chatbot. One reason could be participants' ability to express themselves. We exclusively recruited participants who reported to be fluent in English language, however, their language skills may still have limited their ability to alter their expressions.

Another reason could be that participants actually varied only a small subset of the factors as strategies to trick the chatbot. For example, although the majority of participants identified specific keywords and response time as relevant factors, only a fraction actually employed these strategies. The cause for this mismatch might simply be the difficulty to overview and control all identified factors. Changing the response time while actively concentrating on the content is hard. Also, while participants assumed that keywords or the general tone (e.g., *friendliness*) influence the perception of personality, it is difficult for many to defer which specific keywords could have an impact. To overcome this obstacle, participants invented high-level strategies and, e.g., pretended to be a different person or wrote the opposite of what they had written before.

Since many participants changed the text length of their answers to trick the system, the word count for the second analysis was even lower. Thus, the lack of input words might have led to a smaller divergence from the chatbot's default values for the single traits, which makes the other strategies seem less efficient than they actually are.

Users' Motivation to Trick

Participants agreed that employing tricking strategies is exhausting, tedious and difficult to keep up. Although previous work suggests that online self-presentation is easier to control than implicit behaviour cues in personal interaction [75], the deliberate use of different language requires much concentration even in a written chat scenario. Hence, participants' willingness to trick a system depended on the perceived benefits and consequences. If they expected benefits, e.g., in a job interview, the majority would trick the system to make a better impression. On the other hand, only few participants would actively trick a system to protect their data, particularly if it exceeded one-time interactions. The privacy paradox suggests that users share personal data despite feeling uncomfortable when they expect benefits [10]. Our findings show that this behaviour also holds for *modified* personal data. These findings extend the privacy paradox [10], in which sharing *adapted* data depends on the perceived benefits despite an uncomfortable feeling with sharing personality profiles in general.

However, participants felt a moral obligation not to trick systems or expected negative consequences if other *people* may later associate the tricked profile with them and uncover them as liars. In line with Warshaw et al.'s findings [78], participants found parts of the profile inaccurate but were reluctant to change them, assuming a self-bias. Since participants were not familiar with describing their personality, particularly using the Big Five dimensions, some participants overtrusted the system's profile or even questioned their own self-image.

These findings show a dilemma: Although participants perceived their personality profiles as highly sensitive data, they show reluctance and limited ability to trick a system for automatic personality assessment.

Users' Estimation of the Power of Personality Prediction

Although scandals such as *Cambridge Analytica* made the news worldwide, our participants still underestimated the depth and possibilities of automatic personality assessment, which confirms previous findings [63,64]. They were surprised by how concealed such assessments may happen and showed a very limited awareness of where they could be encountered. Yet, they felt a general discomfort and would prefer to keep such personal information private. However, some participants had already given up on their privacy and felt powerless against the collection of their data.

On the other side, participants overestimated the accuracy and trustworthiness of these systems, which is a known phenomenon in previous research [20]. Many tried to relate to the chatbot profile and played down deviations from their self-image, tending to believe the profile more than themselves. Hence, when users are confronted with an inaccurate profile that is not considered *positive*, particularly less confident or neurotic personalities may start to deeply question themselves, feel embarrassed or even rejected – which can in turn reinforce the underlying cause [25]. We again emphasise that the use of personality assessment systems will need to be carefully pondered, giving weight to peoples' well-being and rights as well as accounting for its limited accuracy.

Design & Social Implications

Our findings indicate that users have difficulties in tricking a personality assessment chatbot even when they are given the possibility to reflect on the interaction and the system's profile output, i.e., to form a mental model and derive strategies to protect themselves. In real life situations, we cannot expect that all systems will offer users feedback about the collected data and inferences – leaving the users unprepared for the quest of protecting their profile and at the mercy of such systems. Even if users are aware of being assessed by such systems, they seem to be overwhelmed by applying strategies in an effective way in the long run. And there certainly are scenarios, such as representing oneself on social media, in which users want to express themselves instead of being perceived as a different or incoherent person. Hence, we argue that HCI researchers as well as society need to protect people and suggest potential steps in multiple directions.

Machine Non-readable Content

Our results as well as Gou et al. [30] showed that users want to protect their data, but not at the cost of leaving a different (and potentially less desired) impression on other users. Participants identified providing less data as an effective strategy to avoid assessment. To allow people to represent themselves freely online, but simultaneously reduce the risk or data for personality assessment, we may think about new possibilities to disguise personal content. For example, we may consider providing systems that present their content readable for humans but not for machines. Textual posts on social media could e.g., be converted to slightly distorted images, which are not readable to machines (c.f. [60]) – similarly to what is already used to grant access to human users and lock out bots.

Educate Users

Our participants were able to identify few effective and practical strategies, although they were given the opportunity to improve their mental model of such algorithms by reviewing and comparing their profiles. The still limited understanding of those systems' functioning and, hence, how they can be tricked shows a need to educate users. To promote self-help, users might be educated about (1) which factors such systems use to assess personality, (2) how these factors and particular expressions are linked to personality characteristics and (3) how to change these expressions to trick systems into capturing a desired profile. For example, a system could analyse texts in real-time and provide the user with feedback and suggestions. In a job interview with a chatbot, e.g., the user may write *'No, I never did that before. I think I learned about this earlier but I am not sure. I would need to read up on it before I would be confident in doing that.'* The tool may explain to the user that this phrase hints at high neuroticism and suggest a more confident reply. While such a system may help users to balance or overcome their weaknesses or understand themselves and their language use better, it may, however, as well be misused.

Inform about System Limitations

Our participants mentioned many times that they would only share the calculated profiles if they could also talk about their limited accuracy and weaknesses. Since personality assessment tools are not yet mature enough to make reliable predictions, as our results show, these tools need to inform all parties – the users who are assessed and the ones collecting the profiles – about the prediction uncertainty and other reliability issues (as also suggested by Lim et al. [45]). Too little or low quality data about the user can cause a low profile accuracy and may lead to a wrong impression and, in turn, to disadvantages for the user. If these systems then predict a future behaviour, e.g., crimes, based on this profile, they may furthermore remind the investigator that there is uncertainty in the profile as well as in the predictions made from it.

Law Enforcement & Moral Obligations

Our participants felt helpless against profiling algorithms and the companies behind them, with the consequence of giving up on their data and privacy. After all, it is questionable whether systems and education alone can actually protect users. We agree with Matz et al. [49] that current legislative approaches are not sufficient since automatic personality assessment can be performed without the user being aware of it. In our opinion, as we cannot account for companies' moral principles, laws have to prohibit the disguise of their data collection and usage terms in endless, non-understandable license agreements and enforce a direct disclosure of such activities together with an opt-in consent. Furthermore, we call for transparency and understandable explanations about the data collection and analysis practices. When users are aware of it, they may be able to act prudently or turn away from it. As for other data, users should be granted access to their profile and be in control of its' use at any time.

SUMMARY & CONCLUSION

In this paper, we investigated whether users can protect themselves from being profiled by a personality assessment chatbot. Our participants interacted twice with the chatbot: They first created their *actual* personality profile by chatting openly with the bot. Participants were then framed with a data privacy and protection story and asked to disguise their personality from the chatbot, i.e., manipulate the system into calculating a profile different from what a non-manipulated assessment would deliver. We showed that every participant could only identify a subset of a large diversity of factors, of which the majority is useful for tricking the system (given its current implementation). Participants were able to trick the chatbot into calculating a slightly different profile by manipulating some of the factors. However, they find the deliberate manipulation too exhausting for long-term use – although they regard personality as very sensitive data and are rather reluctant to share their profiles with others than family and friends.

In addition, we suggest that follow-up research could investigate if users are able to manipulate their profiles in a desired direction and thus assess whether users are able to protect themselves from being profiled *accurately* but also *inaccurately*, depending on the users' goals. In addition, it needs to be understood whether they are able to or can learn to manipulate the single traits to a desired extent, i.e., which strategies apply to which traits to which degree. Furthermore, similar self-protection studies are needed for other forms of interaction, e.g., voice interaction, and other contexts such as social networks and smartphones. Nevertheless, we propose that HCI researchers should provide tools which empower users as well as protect them from being profiled.

Visit the project website for access to the study data files: https://www.medien.ifi.lmu.de/trick-ai

REFERENCES

[1] Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences* 124 (2018), 150–159. DOI: http://dx.doi.org/10.1016/J.PAID.2017.12.018

[2] Jay Belsky, Sara R. Jaffee, Avshalom Caspi, Terrie Moffitt, and Phil A. Silva. 2003. Intergenerational relationships in young adulthood and their life course, mental health, and personality correlates. *Journal of Family Psychology* 17, 4 (2003), 460–471. DOI: http://dx.doi.org/10.1037/0893-3200.17.4.460

[3] David Blackwell, Carrie Leaman, Rose Tramposch, Ciera Osborne, and Miriam Liss. 2017. Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction. *Personality and Individual Differences* 116 (2017), 69 – 72. DOI: http://dx.doi.org/10.1016/j.paid.2017.04.039

[4] Carole Cadwalladr and Emma Graham-Harrison. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. (2018). Retrieved on May 11, 2019 from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

[5] Iván Cantador, Ignacio Fernández-Tobías, and Alejandro Bellogín. 2013. Relating personality types with user preferences in multiple entertainment domains. In *Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization. CEUR Workshop Proceedings*, Shlomo Berkovsky, Eelco Herder, Pasquale Lops, and Olga C. Santos (Eds.). UMAP 2013: Extended Proceedings Late-Breaking Results, Project Papers and Workshop.

[6] Avshalom Caspi, Brent W. Roberts, and Rebecca L. Shiner. 2005. Personality Development: Stability and Change. *Annual Review of Psychology* 56, 1 (2005), 453–484. DOI:http://dx.doi.org/10.1146/annurev.psych.55.090902.141913

[7] Fabio Celli, Elia Bruni, and Bruno Lepri. 2014. Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 1101–1104. DOI:http://dx.doi.org/10.1145/2647868.2654977

[8] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. 2013. Mining Large-Scale Smartphone Data for Personality Studies. *Personal Ubiquitous Comput.* 17, 3 (March 2013), 433 – 450. DOI: http://dx.doi.org/10.1007/s00779-011-0490-1

[9] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. DOI: http://dx.doi.org/10.1177/001316446002000104

[10] Kay Connelly, Ashraf Khalil, and Yong Liu. 2007. Do I do what I say?: Observed versus stated privacy preferences. In *Human-Computer Interaction – INTERACT 2007. Lecture Notes in Computer Science*, C. Baranauskas, P. Palanque, J. Abascal, and S. D. J. Barbosa (Eds.). Springer, Berlin, Heidelberg, 620–623. DOI:http://dx.doi.org/10.1007/978-3-540-74796-3_61

[11] Paul T. Costa and Robert R. McCrae. 1992. Four ways five factors are basic. *Personality and Individual Differences* 13, 6 (1992), 653–665. DOI: http://dx.doi.org/10.1016/0191-8869(92)90236-I

[12] Daniel Danner, Beatrice Rammstedt, Matthias Bluemke, Lisa Treiber, Sabrina Berres, Christopher Soto, and Oliver John. 2016. *Die deutsche Version des Big Five Inventory 2 (BFI-2)*. GESIS - Leibniz-Institut für Sozialwissenschaften, Mannheim, Germany. DOI: http://dx.doi.org/10.6102/zis247

[13] Boele de Raad. 2000. *The Big Five Personality Factors: The psycholexical approach to personality.* Hogrefe & Huber Publishers, Gottingen, Germany.

[14] Colin G. DeYoung. 2014. Openness/Intellect: A dimension of personality reflecting cognitive exploration. In *APA Handbook of Personality and Social Psychology: Personality Processes and Individual Differences*, M. Mikulincer, P.R. Shaver, M.L. Cooper, and R.J. Larsen (Eds.). Vol. 4. American Psychological Association, Washington, DC, USA, 369–399. DOI: http://dx.doi.org/10.1037/14343-017

[15] Ed Diener and Richard E Lucas. 1999. Personality and Subjective Well-being. In *Well-being: Foundations of Hedonic Psychology*, Daniel Kahneman, Ed Diener, and Norbert Schwarz (Eds.). Russell Sage Foundations, New York, NY, USA, 213 – 229.

[16] Ed Diener, Ed Sandvik, William Pavot, and Frank Fujita. 1992. Extraversion and subjective well-being in a US national probability sample. *Journal of Research in Personality* 26, 3 (1992), 205–215. DOI: http://dx.doi.org/10.1016/0092-6566(92)90039-7

[17] M. Brent Donnellan, Rand D. Conger, and Chalandra M. Bryant. 2004. The Big Five and enduring marriages. *Journal of Research in Personality* 38, 5 (2004), 481–504. DOI: http://dx.doi.org/10.1016/j.jrp.2004.01.001

[18] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. 2017. Understanding Shoulder Surfing in the Wild: Stories from Users and Observers. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4254 – 4265. DOI: http://dx.doi.org/10.1145/3025453.3025636

[19] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When People and Algorithms Meet: User-reported Problems in Intelligent Everyday Applications. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA, 96–106. DOI: http://dx.doi.org/10.1145/3301275.3302262

[20] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article Paper 432, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3174006

[21] Anthony M. Evans and William Revelle. 2008. Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality* 42, 6 (2008), 1585–1593. DOI: http://dx.doi.org/10.1016/j.jrp.2008.07.011

[22] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. 2016. Computational personality recognition in social media. *User Modeling and User-Adapted Interaction* 26, 2 (01 Jun 2016), 109–142. DOI: http://dx.doi.org/10.1007/s11257-016-9171-0

[23] Ignacio Fernández-Tobías, Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, and Iván Cantador. 2016. Alleviating the new user problem in collaborative filtering by exploiting personality information. In *User Modeling and User-Adapted Interaction (UMAP '16)*. ACM, New York, NY, USA, 221–255. DOI: http://dx.doi.org/10.1007/s11257-016-9172-z

[24] Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. 2017. Personality Traits and Music Genres: What Do People Prefer to Listen To?. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, New York, NY, USA, 285 – 288. DOI: http://dx.doi.org/10.1145/3079628.3079693

[25] Julie Fitness. 2005. Bye bye, black sheep: the causes and consequences of rejection in family relationships. In *The Social outcast*, Kipling D. Williams, Joseph P. Forgas, and William von Hippel (Eds.). Psychology Press, United States, 263–276.

[26] Rui Gao, Bibo Hao, Shuotian Bai, Lin Li, Ang Li, and Tingshao Zhu. 2013. Improving User Profile with Personality Traits Predicted from Social Media Content. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, New York, NY, USA, 355–358. DOI: http://dx.doi.org/10.1145/2507157.2507219

[27] Alastair J. Gill, Scott Nowson, and Jon Oberlander. 2009. What are they blogging about? Personality, topic and motivation in blogs. In *Third International AAAI Conference on Weblogs and Social Media*. AAAI, Palo Alto, CA, USA, 18–25.

[28] Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting Personality with Social Media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. ACM, New York, NY, USA, 253–262. DOI: http://dx.doi.org/10.1145/1979742.1979614

[29] Lewis R. Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. In *Review of Personality and Social Psychology*, L. Wheeler (Ed.). Vol. 2. Sage Publications, Beverly Hills, CA, USA, 141–166.

[30] Liang Gou, Michelle X. Zhou, and Huahai Yang. 2014. KnowMe and ShareMe: Understanding Automatically Discovered Personality Traits from Social Media and User Sharing Preferences. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 955–964. DOI: http://dx.doi.org/10.1145/2556288.2557398

[31] Guardian. 2018. Amazon ditched AI recruiting tool that favored men for technical jobs. (2018). https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

[32] Gabriella M. Harari, Sandrine R. Müller, Clemens Stachl, Rui Wang, Weichen Wang, Markus Bühner, Peter J. Rentfrow, Andrew T. Campbell, and Samuel D. Gosling. 2019. Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology* Advance online publication (2019). DOI: http://dx.doi.org/10.1037/pspp0000245

[33] Jannica Heinström. 2005. Fast surfing, broad scanning and deep diving: The influence of personality and study approach on students' information-seeking behavior. *Journal of Documentation* 61, 2 (2005), 228–247. DOI: http://dx.doi.org/10.1108/00220410510585205

[34] Michael P. Hengartner, Wolfram Kawohl, Helene Haker, Wulf Rössler, and Vladeta Ajdacic-Gross. 2016. Big Five personality traits may inform public health policy and preventive medicine: Evidence from a cross-sectional and a prospective longitudinal epidemiologic study in a Swiss community. *Journal of Psychosomatic Research* 84 (2016), 44 – 51. DOI: http://dx.doi.org/10.1016/j.jpsychores.2016.03.012

[35] Rong Hu and Pearl Pu. 2010. A study on user perception of personality-based recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization (UMAP 2010)*, P. De Bra, A. Kobsa, and D. Chin (Eds.), Vol. 6075. Springer, Berlin, Heidelberg, 291–302. DOI: http://dx.doi.org/10.1007/978-3-642-13470-8_27

[36] Joshua J. Jackson, Dustin Wood, Tim Bogg, Kate E. Walton, Peter D. Harms, and Brent W. Roberts. 2010. What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality* 44, 4 (2010), 501–511. DOI: http://dx.doi.org/10.1016/j.jrp.2010.06.005

[37] Lauri A. Jensen-Campbell, Ryan Adams, David G. Perry, Katie A. Workman, Janine Q. Furdella, and Susan K. Egan. 2002. Agreeableness, extraversion, and peer relations in early adolescence: Winning friends and deflecting aggression. *Journal of Research in Personality* 36, 3 (2002), 224–251. DOI: http://dx.doi.org/10.1006/jrpe.2002.2348

[38] Lauri A. Jensen-Campbell and William G. Graziano. 2001. Agreeableness as a moderator of interpersonal conflict. *Journal of Personality* 69, 2 (2001), 323–362. DOI:http://dx.doi.org/10.1111/1467-6494.00148

[39] John T. Jost, Jack Glaser, Arie W. Kruglanski, and Frank J. Sulloway. 2003. Political conservatism as motivated social cognition. *Psychological Bulletin* 129, 3 (2003), 339–375. DOI: http://dx.doi.org/10.1037/0033-2909.129.3.339

[40] Timothy A. Judge, Chad A. Higgins, Carl J. Thoresen, and Murray R. Barrick. 1999. The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology* 52, 3 (1999), 621–652. DOI: http://dx.doi.org/10.1111/j.1744-6570.1999.tb00174.x

[41] Meera Komarraju, Steven J. Karau, and Ronald R. Schmeck. 2009. Role of the Big Five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences* 19, 1 (2009), 47–52. DOI: http://dx.doi.org/10.1016/j.lindif.2008.07.001

[42] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805. DOI: http://dx.doi.org/10.1073/pnas.1218772110

[43] Daria J. Kuss and Mark D. Griffiths. 2011. Online Social Networking and Addiction – A Review of the Psychological Literature. *International Journal of Environmental Research and Public Health* 8, 9 (2011), 3528–3552. DOI: http://dx.doi.org/10.3390/ijerph8093528

[44] Lin Li, Ang Li, Bibo Hao, Zengda Guan, and Tingshao Zhu. 2014. Predicting Active Users' Personality Based on Micro-Blogging Behaviors. *PLOS ONE* 9, 1 (01 2014), 1–11. DOI: http://dx.doi.org/10.1371/journal.pone.0084997

[45] Brian Y. Lim and Anind K. Dey. 2011. Design of an Intelligible Mobile Context-aware Application. In *Proceedings of the 2011 International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. ACM, New York, NY, USA, 157–166. DOI: http://dx.doi.org/10.1145/2037373.2037399

[46] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E Moghaddam, and Lyle Ungar. 2016. Analyzing personality through social media profile picture choice. In *Tenth International AAAI Conference on Web and Social Media (ICWSM)*. AAAI, Palo Alto, CA, USA, 211–220.

[47] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* 30 (2007), 457–500.

[48] Gerald Matthews, Ian J Deary, and Martha C Whiteman. 2003. *Personality traits*. Cambridge University Press, Cambridge, UK.

[49] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences* 114, 48 (2017), 12714–12719. DOI:http://dx.doi.org/10.1073/pnas.1710966114

[50] Robert R. McCrae. 2009. The Five-Factor Model of personality traits: consensus and controversy. In *The Cambridge Handbook of Personality Psychology*, Philip J. Corr and Gerald Matthews (Eds.). Cambridge University Press, Cambridge, UK, 148–161. DOI: http://dx.doi.org/10.1017/CBO9780511596544.012

[51] Robert R. McCrae and Paul T. Costa. 2008. A five-factor theory of personality. In *Handbook of Personality: Theory and Research*, O.P. John, R.W. Robins, and L.A. Pervin (Eds.). Vol. 3. The Guilford Press, New York, NY, USA, 159–181.

[52] Robert R. McCrae and Oliver P. John. 1992. An introduction to the five-factor model and its applications. *Journal of Personality* 60, 2 (1992), 175–215. DOI: http://dx.doi.org/10.1111/j.1467-6494.1992.tb00970.x

[53] J. Murray McNiel and William Fleeson. 2006. The causal effects of extraversion on positive affect and neuroticism on negative affect: Manipulating state extraversion and state neuroticism in an experimental approach. *Journal of Research in Personality* 40, 5 (2006), 529–550. DOI: http://dx.doi.org/10.1016/j.jrp.2005.05.003

[54] Matthias R. Mehl, Samuel D. Gosling, and James W. Pennebaker. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology* 90, 5 (2006), 862–877. DOI: http://dx.doi.org/10.1037/0022-3514.90.5.862

[55] Gideon Nave, Juri Minxha, David M. Greenberg, Michal Kosinski, David Stillwell, and Jason Rentfrow. 2018. Musical Preferences Predict Personality: Evidence From Active Listening and Facebook Likes. *Psychological Science* 29, 7 (2018), 1145–1158. DOI: http://dx.doi.org/10.1177/0956797618761659

[56] Scott Nowson. 2007. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *In Proceedings of the AAAI International Conference on Weblogs and Social*. AAAI, Palo Alto, CA, USA.

[57] Maria Augusta S. N. Nunes and Rong Hu. 2012. Personality-Based Recommender Systems: An Overview. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. ACM, New York, NY, USA, 5–6. DOI: http://dx.doi.org/10.1145/2365952.2365957

[58] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, New York, NY, USA.

[59] Daniel J. Ozer and Verónica Benet-Martínez. 2006. Personality and the Prediction of Consequential Outcomes. *Annual Review of Psychology* 57, 1 (2006), 401–421. DOI:http://dx.doi.org/10.1146/annurev.psych.57.102904.190127

[60] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical Black-Box Attacks Against Machine Learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIA CCS '17)*. ACM, New York, NY, USA, 506–519. DOI:http://dx.doi.org/10.1145/3052973.3053009

[61] James W Pennebaker and M E Francis. 1999. *Linguistic Inquiry and Word Count: LIWC*. Erlbaum, Mahwah, NJ, US.

[62] Daniele Quercia, Renaud Lambiotte, David Stillwell, Michal Kosinski, and Jon Crowcroft. 2012. The Personality of Popular Facebook Users. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 955–964. DOI: http://dx.doi.org/10.1145/2145204.2145346

[63] Emilee Rader. 2014. Awareness of Behavioral Tracking and Information Privacy Concern in Facebook and Google. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*. USENIX Association, Menlo Park, CA, USA, 51–67. https://www.usenix.org/conference/soups2014/proceedings/presentation/rader

[64] Ashwini Rao, Florian Schaub, and Norman M. Sadeh. 2015. What do they know about me? Contents and Concerns of Online Behavioral Profiles. *CoRR* abs/1506.01675 (2015), 1–13. http://arxiv.org/abs/1506.01675

[65] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press, Cambridge, UK.

[66] Peter J. Rentfrow and Samuel D. Gosling. 2003. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology* 84, 6 (2003), 1236–1256. DOI: http://dx.doi.org/10.1037/0022-3514.84.6.1236

[67] Giorgio Roffo. 2016. Towards Personality-Aware Recommendation. *CoRR* abs/1607.05088 (2016), 1–4. http://arxiv.org/abs/1607.05088

[68] Matthew Rosenberg, Matthew Confessore, and Carole Cadwalladr. 2018. How Trump Consultants Exploited the Facebook Data of Millions. (2018). Retrieved on May 11, 2019 from https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.

[69] Alexandra Roshchina, John Cardiff, and Paolo Rosso. 2015. TWIN: personality-based intelligent recommender system. *Journal of Intelligent & Fuzzy Systems* 28, 5 (2015), 2059–2071. DOI: http://dx.doi.org/10.3233/IFS-141484

[70] Ramona Schoedel, Quay Au, Sarah Theres Völkel, Florian Lehmann, Daniela Becker, Markus Bühner, Bernd Bischl, Heinrich Hussmann, and Clemens Stachl. 2018. Digital Footprints of Sensation Seeking. *Zeitschrift für Psychologie* 226, 4 (2018), 232–245. DOI: http://dx.doi.org/10.1027/2151-2604/a000342

[71] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, and Martin EP et al. Seligman. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE* 8, 9 (2013), e73791. DOI: http://dx.doi.org/10.1371/journal.pone.0073791

[72] Joseph Sirgy, M. 1985. Using self-congruity and ideal congruity to predict purchase motivation. *Journal of Business Research* 13, 3 (1985), 195–206. DOI: http://dx.doi.org/10.1016/0148-2963(85)90026-8

[73] Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion)*. IEEE, Republic and Canton of Geneva, Switzerland, 107–108. DOI: http://dx.doi.org/10.1145/2872518.2889368

[74] Clemens Stachl, Quay Au, Ramona Schoedel, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, Heinrich Hussmann, Bernd Bischl, and Markus Bühner. 2019. Behavioral Patterns in Smartphone Usage Predict Big Five Personality Traits. *PsyArXiv* (2019), 1–24. DOI: http://dx.doi.org/10.31234/osf.io/ks4vd

[75] Michele M. Strano. 2008. User Descriptions and Interpretations of Self-Presentation through Facebook Profile Images. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 2, 2 (2008). https://cyberpsychology.eu/article/view/4212

[76] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing* 5, 3 (2014), 273–291. DOI: http://dx.doi.org/10.1109/TAFFC.2014.2330816

[77] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Quay Au, Bernd Bischl, Markus Bühner, and Heinrich Hussmann. 2019. Opportunities and Challenges of Utilizing Personality Traits for Personalization in HCI: Towards a shared perspective from HCI and Psychology. In *Personalized Human-Computer Interaction*, Mirjam Augstein, Eelco Herder, and Wolfgang Wörndl (Eds.). De Gruyter, Oldenbourg, Germany, Chapter 2, 31–66. DOI: http://dx.doi.org/10.1515/9783110552485-002

[78] Jeffrey Warshaw, Tara Matthews, Steve Whittaker, Chris Kau, Mateo Bengualid, and Barton A. Smith. 2015. Can an Algorithm Know the "Real You"?: Understanding People's Reactions to Hyper-personal Analytics Systems. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 797–806. DOI: http://dx.doi.org/10.1145/2702123.2702274

[79] Robert E. Wilson, Samuel D. Gosling, and Lindsay T. Graham. 2012. A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science* 7, 3 (2012), 203–220. DOI: http://dx.doi.org/10.1177/1745691612442904 PMID: 26168459.

[80] Tal Yarkoni. 2010. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality* 44, 3 (2010), 363–373. DOI: http://dx.doi.org/10.1016/j.jrp.2010.04.001

[81] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (2015), 1036–1040. DOI:http://dx.doi.org/10.1073/pnas.1418680112

[82] Michelle X. Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting Virtual Agents: The Effect of Personality. *ACM Trans. Interact. Intell. Syst.* 9, 2-3, Article 10 (March 2019), 36 pages. DOI: http://dx.doi.org/10.1145/3232077